

**Comparative Genomics of Microbial Chemoreceptor
Sequence, Structure, and Function**

A Dissertation Presented for the
Doctor of Philosophy
Degree
The University of Tennessee, Knoxville

Aaron Daniel Fleetwood
December 2014

Copyright © by Aaron Daniel Fleetwood
All rights reserved.

Dedication

To my Wife,

Dr. Ellen Ann Fleetwood

To my Family,

Dan, Betsy, Zach, and Nathan

To The Messengers,

Mark, Theresa, and Paul

And

To Friends

Cougar, Ketra, Nash, Ginger, and Piece

Acknowledgments

I would like to express sincere and profound appreciation to my mentor, Dr. Igor B. Zhulin, who has been an unwavering source of support and guidance in addition to expanding and sharpening my scientific worldview. I would also like to thank the Genome Science and Technology Program for taking a chance on what on paper might have appeared to be an economist and linguist, not a biologist. In particular, Dr. Albrecht von Arnim has been a tremendous resource and advisor over the years. I am ever grateful to my committee members, Dr. Robert Hettich, Dr. Elias Fernandez, and Dr. Elizabeth Howell. They, along with Dr. Zhulin, are responsible for fostering and nurturing my passion for genome and protein science, as well as shaping my scientific philosophy at every stage and level of my graduate experience. Additionally, many outstanding faculty, including Dr. Jerome Baudry and Dr. Tim Sparer, were instrumental in both my academic education and my growth as a scientist and professional. Thank you also to my lab mates Amit Uphadyay, Ogun Adebali, Dr. Kirill Borziak, and Dr. Davi Ortega for wonderful meetings, discussions, and camaraderie. I also owe a debt of gratitude to past members of the Zhulin Lab whose past and current work have served as a foundation and inspiration for my research (Drs. Kristen Wuichet, Luke Ulrich, Roger Alexander, and Brian Cantwell). Finally, thank you to our experimental collaborators and advisors (Drs. Brian Crane, Victoria Korolik, Caroline Harwood, and Harry Mobley) for bringing our ideas to *vivo*.

I would not be here today without tremendous support and funding from the Genome Science and Technology Program, the J. Wallace and Katie Dean Multi-Year Graduate Fellowship, the UTK Graduate School, the Microbiology Department, Oak Ridge National Laboratory and the National Institutes of Health.

Abstract

Microbial chemotaxis receptors (chemoreceptors) are complex proteins that sense the external environment and signal for flagella-mediated motility, serving as the GPS of the cell. In order to sense a myriad of physicochemical signals and adapt to diverse environmental niches, sensory regions of chemoreceptors are frenetically duplicated, mutated, or lost. Conversely, the chemoreceptor signaling region is a highly conserved protein domain. Extreme conservation of this domain is necessary because it determines very specific helical secondary, tertiary, and quaternary structures of the protein while simultaneously choreographing a network of interactions with the adaptor protein CheW and the histidine kinase CheA. This dichotomous nature has split the chemoreceptor community into two major camps, studying *either* an organism's sensory capabilities and physiology *or* the molecular signal transduction mechanism. Fortunately, the current vast wealth of sequencing data has enabled comparative study of chemoreceptors. Comparative genomics can serve as a bridge between these communities, connecting sequence, structure, and function through comprehensive studies on scales ranging from minute and molecular to global and ecological. Herein are four works in which comparative genomics illuminates unanswered questions across the broad chemoreceptor landscape. First, we used evolutionary histories to refine chemoreceptor interactions in *Thermotoga maritima*, pairing phylogenetics with x-ray crystallography. Next, we uncovered the origin of a unique chemoreceptor, isolated only from hypervirulent strains of *Campylobacter jejuni*, by comparing chemoreceptor signaling *and* sensory regions from *Campylobacter* and *Helicobacter*. We then selected the opportunistic human pathogen *Pseudomonas aeruginosa* to address the question of assigning multiple chemoreceptors to multiple chemotaxis pathways within the same organism. We assigned all *P. aeruginosa* receptors to pathways using a novel *in silico* approach by incorporating sequence information spanning the entire taxonomic order *Pseudomonadales* and beyond. Finally, we surveyed the chemotaxis systems of all environmental, commensal,

laboratory, and pathogenic strains of the ubiquitous *Escherichia coli*, where we discovered an ancestral chemoreceptor gene loss event that may have predisposed a well-studied subpopulation to adopt extra-intestinal pathogenic lifestyles. Overall, comparative genomics is a cutting edge method for comprehensive chemoreceptor study that is poised to promote synergy within and expand the significance of the chemoreceptor field.

Preface

In the following dissertation, comparative genomics methods were used to study microbial chemoreceptor sequences, structures, and functions. Microbial in this case refers primarily to *Bacteria*, though chemoreceptors are found in *Archaea* as well (hence the broader term in the title). The first chapter will be an introduction, in which I provide a brief overview of the motivations behind this work before diving into the scientific discipline (computational biology), methods (comparative genomics), and tools (bioinformatics) that are utilized throughout. This section covers aspects ranging from the basic (core tenets of biology) to the technical (pitfalls and nuances of comparative work). I will then introduce the field of chemotaxis with a special emphasis on chemoreceptors. This introductory “review” will be further subdivided into two major sections corresponding to the two major communities that study chemoreceptors: structure-focused signal transduction work and microbiology-based behavioral/physiological studies. The motivations of these two fields are quite different, with the former pursuing more fundamental and mechanistic understanding and the latter seeking applied connections, such as links to pathogenicity.

After the introductory chapter, the main body of the dissertation is divided into four additional chapters which correspond to full peer-reviewed publications, manuscripts submitted for review, or featured aspects of manuscripts in preparation, all of which have been produced during my graduate work. Biology today (especially computational work) is a collaborative and highly interdisciplinary endeavor, so for each chapter I provide the contributions of myself and the other authors, as well as contributions outside of authorship but still deserving of mention. Each chapter will consist of an abstract, introduction, results, and discussion section, and any materials and methods specific to that paper or published in the manuscript will accompany these chapters as well. Chapter 3 (Assigning Chemoreceptors to Pathways in *Pseudomonas aeruginosa*) represents my most significant independent effort as a graduate student. Beginning

with the first chapter, the studies increase in organizational complexity from studying a handful of proteins in a limited set of organisms to comparing entire sets of chemotaxis systems across broad taxonomic divisions. This serves to illustrate the power of comparative genomics to adapt to the scale of the biological problem, and for proteins like chemoreceptors, this is a tremendous methodological advantage.

After the main body, the concluding chapter will contain my opinions on where the chemotaxis field may be heading in the future, especially if comparative genomics can be used to tie sequence and structural work to physiological observations and functions. This dissertation provides evidence that this type of interdisciplinary synergy is not only possible, but has already yielded significant contributions to the chemoreceptor field. I will also present specific suggestions for future work that can be used to extend the impact of this dissertation in years to come.

Table of Contents

Chapter 1: Introduction	1
Overview	1
Comparative Genomics Materials and Methods	3
Chemoreceptors and Chemotaxis: A House Divided	27
Signal Transduction	31
Microbial Behavior and Pathogenicity	35
Chapter 2: Refining Chemoreceptor Interactions in <i>Thermotoga maritima</i>	45
Chapter Source and Author Contributions	46
Abstract	47
Introduction	48
Results	53
Discussion	68
Chapter 3: Investigating a Novel Chemoreceptor in <i>Campylobacter jejuni</i>	77
Chapter Source and Author Contributions	78
Abstract	79
Introduction	80
Results	81
Discussion	87
Chapter 4: Assigning Chemoreceptors to Pathways in <i>Pseudomonas aeruginosa</i>	91

Chapter Source and Author Contributions	92
Abstract	93
Introduction	94
Results	98
Discussion	109
Chapter 5: Chemoreceptor Gene Loss and Pathogenicity in <i>Escherichia coli</i>	113
Chapter Source and Author Contributions	114
Abstract	115
Introduction	116
Results	119
Discussion	127
Chapter 6: Conclusions	131
List of References	138
Appendices	154
Appendix A	155
Appendix B	156
Vita	157

List of Tables

1. Data Collection and Refinement Statistics for Ternary Complex Crystal Structure.....	58
2. Pfam Domain Results for <i>Campylobacter</i> and <i>Helicobacter</i> Chemoreceptors.....	81
3. HHpred Domain Results for <i>Campylobacter</i> and <i>Helicobacter</i> Chemoreceptors.....	82
4. Chemotaxis Systems, System Designations, and Heptad Class Associations.....	100
5. Loss of Chemoreceptor Genes in <i>E. coli</i> and <i>Shigella</i> Genomes.....	123
6. Horizontally Transferred Chemoreceptor Genes in <i>Escherichia</i> Genomes.....	125

List of Figures

1. Protein Domains Can Evolve at Different Rates (Chemoreceptor Example).....	9
2. Evolutionary History is Key to Productive Comparative Analysis.....	12
3. Sample Multiple Sequence Alignments and Visualization Strategies.....	16
4. Overview of Producing a Profile HMM and Identifying Matching Sequences.....	23
5. Diagram of Canonical and Non-Canonical Chemoreceptor Localization.....	32
6. Established Roles of Chemotaxis Systems Contributing to Pathogenicity in Humans.....	38
7. Molecular Composition of the 3.2 Å Resolution Ternary Complex Crystals.....	54
8. Unzipping of the Tm14 _S Helical Hairpin.....	56
9. Pseudosymmetric Contacts Made by the CheW and P5 Subdomains.....	60
10. Interaction Between Tm14 _S and P5 or CheW.....	61
11. Phylogenetic Trees Showing Two Groups of CheW Orthologs from <i>Thermotogae</i>	64
12. Structure—Function Analysis of Ternary Complex Interfaces.....	71
13. Hypothetical Alternate Interfaces in Chemosensory Clusters.....	75
14. Pairwise Comparisons of CcrG to Aspartate and Multi-Ligand Receptors in <i>C. jejuni</i>	83
15. Maximum Likelihood Tree of Campylobacter and Helicobacter Chemoreceptors.....	84
16. Sequence and 2D Structural Alignment of Cache_1 Chemoreceptors.....	86
17. Homology Modeling of LBD with Implications for Amino Acid Specificity Change.....	89
18. <i>A Priori</i> Chemotaxis Systems and Chemoreceptors in <i>Pseudomonas aeruginosa</i>	99
19. Helical Heptad Distribution of Chemoreceptors in PAO1 (26 MCPs).....	101
20. Helical Heptad Length of Chemoreceptors in 1 System Organisms.....	102
21. 16S RNA Tree of Order <i>Pseudomonadales</i> with Chemotaxis Phylogenetic Profiles.....	104
22. 696 Chemoreceptor Signaling Regions from <i>Pseudomonadales</i>	108
23. New Look: Chemotaxis Systems PAO1 Disentangled.....	111
24. Presence of Chemotaxis Genes in Completely Sequenced <i>Escherichia/Shigella</i>	122

25. Deletions in <i>tap</i> and <i>trg</i> Genes in B2 Group Strains.....	124
26. Filling in the Gaps of Chemoreceptor Sensory Capabilities.....	133
27. Convergent Evolution of Soluble Metabolism-Sensing Chemoreceptors?.....	136

List of Abbreviations

1. MCP	Methyl-accepting Chemotaxis Protein (Chemoreceptors)
2. HCD	Highly Conserved Domain
3. MCPsignal	Methyl-accepting Chemotaxis Protein Signaling Domain
4. LBD	Ligand Binding Domain
5. <i>tsr</i>	Taxis toward Serine Receptor
6. <i>tar</i>	Taxis toward Aspartate Receptor
7. <i>trg</i>	Taxis toward Ribose and Glucose
8. <i>tap</i>	(Dipeptide and Pyrimidine Chemotaxis Transducer)
9. <i>aer</i>	Aerotaxis receptor
10. <i>pct(ABC)</i>	<i>Pseudomonas</i> Chemotaxis Transducer (A, B, and C paralogs)
11. TM	Transmembrane
12. Che_	Chemotaxis protein _ (I.e. CheA is Chemotaxis protein A)
13. Tlp	Transducer-like protein (often used for chemoreceptors as well)
14. N	Amino terminus of a protein (see below for amino acid)
15. C	Carboxyl terminus of a protein
16. MSA	Multiple Sequence Alignment
17. HMM	Hidden Markov Model
18. PSSM	Position-Specific Scoring Matrix
19. HAMP	Histidine kinases, Adenylyl cyclases, Mcps, and some Phosphatases
20. HPT	Histidine containing-PhosphoTransfer
21. Tm14 _s	<i>Thermotoga maritima</i> Protein Interaction Region (TM0014)
22. MCP _C	MCP Cytoplasmic region
23. KCM	Kinase Control Module
24. PIR	Protein Interaction Region

25. rRes	Receptor Residue (e.g. rArg150)
26. kRes	Kinase Residue (e.g. kArg150)
27. wRes	CheW Residue (e.g. wArg150)
28. MR	Molecular Replacement
29. ECT	Electron CryoTomography
30. PDS	Protein Dipolar electron-spin resonance Spectroscopy
31. TAM-IDS	Tryptophan and Alanine Mutation to Identify Docking Sites
32. CcmL	<i>Campylobacter</i> Chemoreceptor for Multiple Ligands
33. CcaA	<i>Campylobacter</i> Chemoreceptor for Aspartate A
34. CcrG	<i>Campylobacter</i> Chemoreceptor for Galactose
35. TFP	Type-Four Pilus (Chemotaxis System Subset)
36. ACF	Alternative Cellular Function (Chemotaxis System Subset)
37. WSP	Wrinkly Spreader Phenotype
38. CHP	CHemosensory Pili system
39. spp.	Species (plural)
40. EPEC	EnteroPathogenic <i>E. coli</i>
41. Ex-PEC	Extra-intestinal Pathogenic <i>E. coli</i>
42. APEC	Avian Pathogenic <i>E. coli</i>
43. MNEC	Meningitis-associated Neonatal pathogenic <i>E. coli</i>
44. UPEC	Urinary Pathogenic <i>E. coli</i>
45. sSNP	synonymous Single Nucleotide Polymorphism
46. Ala (A)	Alanine
47. Arg (R)	Arginine
48. Asn (N)	Asparagine
49. Asp (D)	Aspartic Acid
50. Cys (C)	Cysteine

51. Glu (E)	Glutamate
52. Gln (Q)	Glutamine
53. Gly (G)	Glycine
54. His (H)	Histidine
55. Ile (I)	Isoleucine
56. Leu (L)	Leucine
57. Lys (K)	Lysine
58. Met (M)	Methionine
59. Phe (F)	Phenylalanine
60. Pro (P)	Proline
61. Ser (S)	Serine
62. Thr (T)	Threonine
63. Trp (W)	Tryptophan
64. Tyr (Y)	Tyrosine
65. Val (V)	Valine

*11-30 correspond to the 20 common amino acids. Many figures (especially multiple sequence alignments) utilize the single letter abbreviation scheme.

List of Attachments

1. Supplement for 2013 *Biochemistry* Paper (Ch. 2).....Appendix A.pdf
2. Supplement for 2013 *Journal of Bacteriology* Paper (Ch. 5).....Appendix B.pdf
3. Supplementary Dataset for 2013 *Journal of Bacteriology* Paper (Ch. 5).....Dataset B1.xlsx

List of Symbols

1. Å Angstrom (Ångström)
2. α Alpha (as in α -helix, or α -*Proteobacteria*)
3. β Beta (as in β -strand or sheet, or β -*Proteobacteria*)
4. δ Delta (δ -*Proteobacteria*)
5. ϵ Epsilon (ϵ -*Proteobacteria*)
6. γ Gamma (γ -*Proteobacteria*)
7. ϕ Phi (as in ϕ -angle)
8. ψ Psi (as in ψ -angle)
9. ΔG Change in Gibbs Free Energy
10. Σ Sigma

Chapter 1: Introduction

Overview

Three major sources have influenced this dissertation: a rapidly evolving technology (genome sequencing), an innovative theory-driven methodology (comparative genomics), and a time-tested model system (bacterial chemotaxis). Work like this was not possible even 10 years ago, as it relies heavily on great strides made in genome sequencing technology to produce sequences for comparative analysis. In order to conduct said analysis, a combination of comparative genomics and protein sequence analysis provides the most direct route to connecting raw sequences to biological reality. Finally, cutting edge data and methods require a well-studied system to probe, and bacterial chemotaxis is one of the most extensively characterized systems in all of biology. While genome sequencing technology made this work possible, the latter two elements will be the focus of this dissertation.

The “genomic age” for biology began with the initial sequencing and analysis of the human genome in 2001.^{1,2} The technological capabilities for sequencing genomes has advanced at beyond exponential rates, and the resulting eruption of data has been staggering. The sheer amount of sequence data is a technical challenge unto itself, but taking raw sequence data and connecting it to assayable biology is a different matter entirely. The DNA being sequenced in these genomes contain coding regions for genes, which in turn can be translated into protein sequences of amino acids. Proteins are the prime effectors of biology, serving crucial roles as both the structure and the molecular machines for all living cells. By knowing a protein sequence, we effectively have a blueprint that provides clues as to how that protein may function. Understanding how proteins function is fundamental to our characterization of well-studied and novel biological systems. Moreover, many diseases have been connected to protein function (or dysfunction), making them highly actionable avenues for therapeutic study.

Protein coding sequences are rarely held constant; rather, these sequences are major sources of evolutionary adaptation within all life. Essential or beneficial genes are passed on to the next generation, and during this process, mutations can occur. Mutations may be neutral, exquisitely fine-tune the existing function, or dramatically alter and even break the protein for which they code. However, there are positions in sequences that resist the urge to change, in some cases staying constant across vast evolutionary distances. These positions are conserved, and one of the core tenets of biology is that conserved positions in proteins are essential for their structure and function. Thus, by studying how proteins change over time, we gain insight into how they work in the present. Furthermore, by comparing how the sequences that encode proteins exist in a genomic context, we gain expanded insight on both that protein's function and its potential contribution to observable behavior.

As for the model system, bacterial chemotaxis is one of the most highly characterized systems, where the majority of involved proteins have known functions and assayable phenotypes (allowing for productive comparative genomics analysis). Chemotaxis is the phenomenon by which motile bacteria with flagella navigate through environments to find those which are suitable for their survival and growth. Chemoreceptors are the proteins in this system that sense environmental signals such as nutrients, toxins, or compounds from other bacteria, “steering” movement toward attractants and away from repellents. Because of this crucial role, the number and type of chemoreceptors found in a given organism (the chemoreceptor suite or repertoire) serve like a GPS (Global Positioning System). There are several proteins that directly interact with chemoreceptors which, along with a multitude of accessory and regulatory proteins, provide numerous functional contributions and interactions which can be probed both experimentally and computationally. That said, there are still many unanswered questions remaining for chemotaxis despite its intensive study, and chemoreceptors tend to be at the center of them all.

For the overall structure of the dissertation, we begin with a comprehensive introduction to comparative genomics methods including theories, assumptions and justifications that shape the overarching scientific philosophy guiding the work. Afterwards, we provide a thorough review of chemotaxis from two major viewpoints: first as a mechanistic model system for structural biologists, and secondly as a biological system that contributes to overall behavior of microbes ranging from commensal organisms to major pathogens. Next, the main body of the dissertation comprises four chapters that span a broad range of biological scales, demonstrating how the combination of structural information and phenotypic information can potentially provide more biological relevance for the study of chemoreceptors than either alone. Finally, we conclude with predictions for where the study of chemoreceptors is heading, followed by concrete suggestions for future work that may contribute to actualizing these forecasts.

Comparative Genomics Materials and Methods

While genes have been inherited and have been mutating since the advent of life, traces of these events are visible today. There are patterns and tendencies for genetic change that can be observed, analyzed, and probed experimentally. One example for this is that when an evolutionary innovation works, it may confer survival advantages and propagate across a wide diversity of organisms. Further iterations of this gene become variations on a theme, diverging away from a common ancestral gene. These variations may sample a wide space of possibilities, but there are natural forces (e.g. chemistry and physics) that shape and constrain this space so that we can detect similarities between related sequences. This makes comparative approaches extremely powerful methods of biological inquiry when the right tools and perspective meet the right sequences. With the maturation of genome sequencing technologies and ever-increasing computational power, it is both exhilarating and also an extreme privilege to study biology in this era.

Computational biology is a large scientific discipline possessing a tremendous range of diverse methodologies. As the name implies, the commonality is working *in silico* rather than in the wet lab. At the core, computational biologists use bioinformatics methods and tools to interrogate biological questions. Computational biology is often confused with bioinformatics, but the distinction is that bioinformatics focuses on method and tool development rather than biological analysis (though members from each group often cross this line). This discipline is complementary to traditional experimental methodologies, and can generate and test hypotheses as well as inform and guide other experimental disciplines. To be successful, one must be able to critically assess bioinformatics tools in addition to providing justifications of assumptions and controls, in much the same fashion as traditional experimental work. As such, one cannot overlook the necessities of appreciating the work from both communities and fostering collaborative interdisciplinary relationships.

From a computational standpoint, one can study proteins over a range of biological scales from the molecular and structural level (biochemistry and biophysics) all the way up to the physiological behavior of an organism (microbiology and ecology). When a protein is placed within a greater biological context, one may be able to decipher how a given protein interacts with other proteins and systems in an organism (systems biology). Doing so often requires an understanding of the evolutionary history of the gene encoding that protein (phylogenetics), which provides fundamental insight that no other experimental methodology is able to replicate. In order to accomplish all of these goals, one can investigate how protein sequences (and the genes encoding them) change across different organisms (comparative genomics). Comparative genomics, by a broad definition, takes sequence data and uses studied and characterized examples or well-established biological theory to explain, predict, or advance our understanding of comparable unknowns.

This work is guided by several core biological principles and theories. The first and foremost of these is that the amino acid sequence of a protein determines its structure, which in turn determines its function. Even in the case of unstructured or disordered proteins, they owe that very disorder and its functional role to the sequence. Second, despite this fundamental understanding, working with only a sequence and arriving at the function of a protein is rarely a straightforward process (unless the sequences are almost identical). In doing comparative work, one cannot infer function without having witnessed or connected this to previous studies or experience. Even so, it stands as a powerful method to answer unanswered biological questions, including refining and enhancing previous conceptions of even well-established functions. Connecting sequence to function in the most direct, accurate, and reproducible manner is a major fundamental question facing biology and is a recurring theme throughout this dissertation. This is especially significant in light of the flood of sequences and relative lack of characterized structures and functions.

Therefore, to leverage the vast array of sequences in the most efficient way possible, comparative genomics with a focus on protein sequence analysis is the experimental methodology of choice to answer this question. In all cases, the most important aspect of comparative genomics is choosing the right comparative target for one's query. To this end, there are very few established protocols to accomplish this initial and critical step, and a great deal of research and understanding of the relationships between organisms (taxonomy) is essential. Furthermore, a firm command of evolutionary events and the potential trajectories and fates of protein-coding genes is also invaluable. When one does produce a comparative set, it is also important to establish what type of relationship exists between the two groups. Often in biology, the term homology is used when one protein is similar to another. However, there are multiple levels and classes of homologous relationships, each with potential ramifications for the

results of any analyses that would result from a comparison. The following section addresses these distinctions in greater detail.

Basic Guiding Theories and Concepts. The central dogma of biology is that genetic information is encoded and transmitted to the next generation through DNA, which is transcribed into RNA, which is then decoded into proteins via translation. Proteins are molecular machines and effectors of biology, giving life shape, form and function. The amino acid sequences encoded by DNA gives rise to protein structure, and another core tenet of biological theory is that the structure of a protein determines its function. However, ascribing a function to a sequence is anything but straightforward, and there are multiple strategies and levels of analysis involved in doing so.

Part of these challenges stem from the fact that proteins have multiple levels of structural complexity. These levels are products of both the protein sequence and the environment in which they are expressed.³ The first of these, primary structure, is the sequence of amino acids that is the result of ribosomal translation and is the basis for protein sequence analysis. Amino acids are small molecules that share an amino group and a carboxyl group, but are defined by a third component: the R (functional) group. The R group determines the biochemical properties of that particular amino acid, which in turn influence the formation of local structure, catalytic activity, and many other characteristics once it becomes part of a protein sequence. There are 20 common amino acids (see List of Abbreviations), and the relationships of these to one another form much of the qualitative side of protein sequence analysis.

During the formation of the primary structure of a protein, amino acids are covalently linked by peptide bonds, losing either a hydrogen atom from their amino group and/or a hydroxyl group from their carboxyl group.⁴ Once incorporated into a polypeptide chain, amino acids are referred to as amino acid residues (or simply residues). The residue at the end of the

polypeptide that retains a full amino group becomes the N-terminus and beginning of the protein, while the residue that retains a full carboxyl group becomes the C-terminus and end of the protein. Both during and after translation of the primary structure, secondary structure results from intramolecular associations between amino acid residues that form defined local structures via hydrogen bonding such as alpha helices and beta strands. Regions that lack this defined structure can form loops of variable lengths, and these can serve no known function, serve as functional linkers, or even contribute to catalytic activity that is the main function of the overall protein. The next level, tertiary structure, begins to take shape when the hydrophobic core of the protein and the elements of secondary structure begin to fold the protein into a distinct shape, generally referred to as a globular domain. This is not always the case, and some proteins functionally resist forming higher order structures and are intrinsically disordered. Finally, quaternary structure forms when two or more proteins associate, and can involve the pairing of two identical proteins (homodimers), two distinct proteins (heterodimers), or multiplicities of interactions (oligomerization or protein complex formation). In summary, each of these levels of structure can contribute to the overall physiological functioning of any given protein.

Protein Domains. The protein domain is a unit of a protein that can fold and function independently of the rest of the protein. The average protein domain is 100 amino acids in length,⁵ and the average protein contains 2-3 domains.⁶ The domain composition of a protein is often referred to as its domain architecture. Since protein domains can operate “semi-autonomously”, domains within the same protein can experience different evolutionary forces. In the chemoreceptor, for example, the ligand binding domain (LBD or sensory region) is exposed to the extracellular environment and must adapt to diverse environmental conditions. In the same protein, the highly conserved signaling domain must maintain multiple simultaneous interactions, requiring the right residue in the right position for the vast majority of

chemoreceptor space (**Figure 1**). Even when a domain acquires a new function (like changing its sensory specificity), that domain can be shuffled with other domains to make an entirely different protein under the right circumstances. Using this information, it is possible and often essential to work with “pieces” of proteins individually, so one of the first and foremost skills for a protein sequence analyst is to identify domains, their borders, and their contribution to the overall function of a given protein.

Evolutionary History, Homology, and Types of Homologs. There are several major mechanisms that can drive evolution. However, in the simplest of cases, a parent passes on a gene to their offspring, resulting in vertical evolution. When this occurs, that gene will encode a similar protein with the same structure, domain architecture and function as it did in the previous generation. There are several events that can cloud these straightforward 1:1 relationships, and these events are also major drivers of evolution. Major examples of these events are gene loss, gene duplication, and horizontal gene transfer (explanations to follow). In all cases, genetic changes can have a dramatic effect on a protein. These events serve to alter or test the evolutionary pressure, resulting in selection. While gene loss is difficult to observe, gene duplication and horizontal transfer are fairly straightforward to identify with the correct methods. Despite relatively simpler detection, their presence can confound results, posing added challenge when evolution is not purely vertical.

Most biologists who are not primarily concerned with evolutionary relationships or histories of proteins use the term **homolog** when referring to any similar version of a protein encoded within a different organism. Homolog stems from the term homology, which is a measure of similarity or relatedness between two proteins. For instance, there are five chemoreceptors present in *Escherichia coli*, so there are five chemoreceptor homologs. However, there is no universal agreement on how to measure similarity in quantitative terms, so the percentage of identical amino acids shared in the same position in each sequence is often

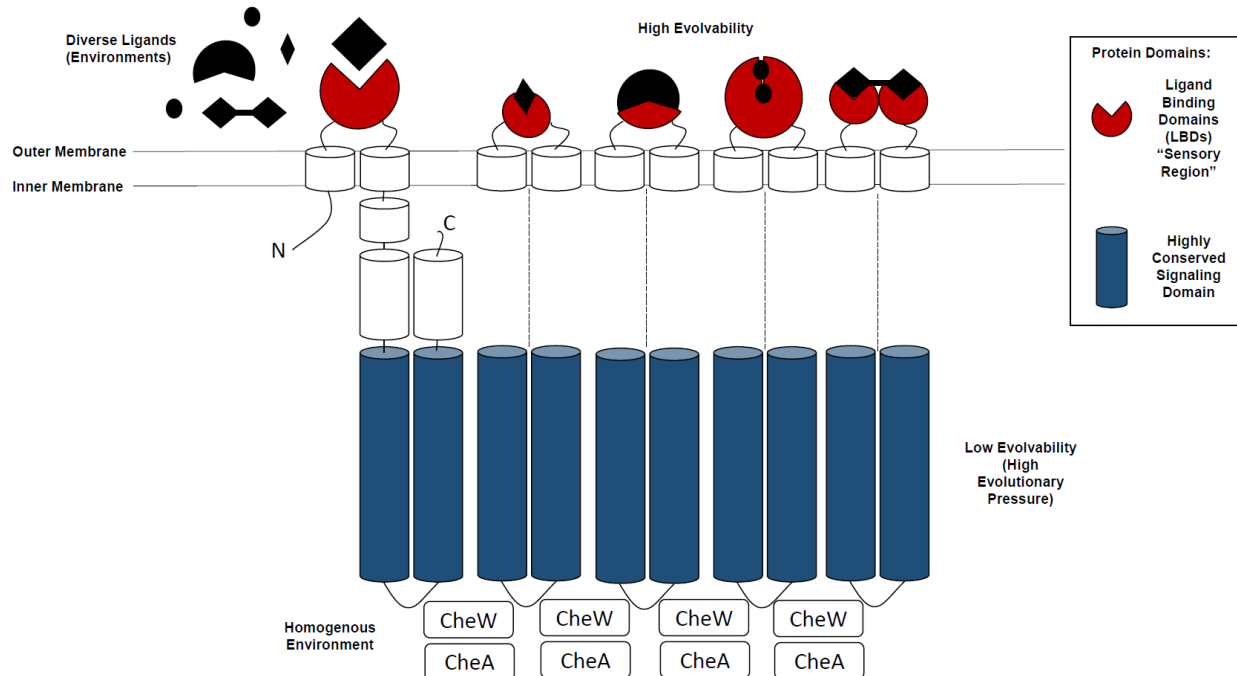


Figure 1. Protein Domains Can Evolve at Different Rates. This diagram shows a chemoreceptor localized to the membrane (far left, bounded by the N and C termini) with two major domains highlighted and architecturally represented (red: LBD, blue: HCD). Four other chemoreceptors are shown to the right (simplified for illustrative purposes), and these have similar signaling domains due to the homogenous intracellular environment of interacting with the proteins CheW and CheA. However, these receptors can contain radically different sensory domains (both from adapting an existing domain or using an entirely unrelated domain), reflecting the diversity of the environment to which this organism has adapted. Comparison of the HCD would be fairly straightforward, whereas comparison of the LBDs may not be possible at all.

used. There are conventions that determine cutoffs based on percentage identity for homology, but the reliability of using generalized cutoffs varies greatly with the nature of the protein. Other schema consist of using similarities in biochemical properties of residues to bolster relatedness claims, but again, there is no consensus on quantifying how similar two protein sequences are to one another. Thus, conventions were needed to establish different types of homologs, based

upon their evolutionary history and genomic context, in order to more reliably and consistently compare related sequences.

For our purposes, comparative genomics employs terminology to make distinctions when comparing protein sequences.⁷ Homolog, in this discourse, is a broad all-encompassing term that must be subdivided into particular classes of homologs: **orthologs**, **paralogs**, and **xenologs**. As stated previously, vertical evolution is the simplest fate for any given gene, producing homologs that are often almost identical. These cases are termed **orthologs**, and comparison within this group is relatively straightforward. Both *Salmonella* and *Escherichia* species contain an aspartate-sensing chemoreceptor that has evolved vertically, making the two proteins orthologs. However, when a gene is duplicated and two copies are transmitted to the next generation, those homologs will now experience very different scenarios. Leading evolutionary theories suggest that one possible outcome for gene duplication is that the presence of two genes relaxes evolutionary pressure on each, allowing them to mutate and explore functional space. Normally, one gene will continue to perform the original function (especially when the function is necessary for the survival of the organism), while the other gene may acquire a new but related function (*neofunctionalization*).⁸ These two homologs are now **paralogs**. If that new function confers a survival advantage, it will be fixed through increased evolutionary pressure to maintain the new competitive advantage. Paralogs are not limited to single duplication events: chemoreceptors *pctA*, *pctB*, and *pctC* from *Pseudomonas aeruginosa* arose from two duplication events and each senses a different set of amino acids (grouped by biochemical properties). Often, these changes necessitate the removal of such sequences from datasets in order to maintain assumptions of like evolutionary pressure.

To further complicate evolutionary history, genetic information can be exchanged between organisms (especially bacteria) through HGT (horizontal gene transfer, or lateral gene transfer). Bacteria are particularly adept at exchanging genes through competence systems,

plasmid transfer, and bacterial conjugation and will often transfer beneficial genes throughout a microbial community. In these cases, the gene transferred will look much more like genes from the transferring organism than the new host, though this effect will diminish over time as the gene takes on the coding and sequencing preferences of the new host. This type of homolog, now a **xenolog**, can outperform and replace a similar gene (*xenologous gene displacement*). Because of xenologous gene displacement and the aforementioned evolutionary twists and turns of paralogy, we must carefully assess the evolutionary history of the proteins that we compare before we begin to analyze them, so that we can assure that we are making the most direct and biologically relevant comparisons (**Figure 2**).

Taxonomy, Phylogenetics, and Evolutionary Distance. Scientists and naturalists have attempted to impose classification schemes (taxonomy) on living organisms since the advent of biology. While earlier systems were based on physical observations or behavioral characteristics, more sophisticated molecular mechanisms have taken their place. Woese and Fox's 16s RNA classification took advantage of the extreme conservation and subtle variations in ribosomal subunits in order to build the first genetically based taxonomic system.⁹ Many other methods have arisen since, including concatenation (stringing together) of numerous conserved housekeeping genes, and even whole genome comparisons. The guiding principle behind these systems is that the more divergent the sequence, the more time has elapsed between the speciation of any two given organisms, as more mutations (which are temporally rare events) have accrued. The major caveats to this assumption are that the more important the gene, the less prone it will be to mutation, and also that horizontally transferred genes can cross between distantly related organisms (uncoupling them from any meaningful temporal comparison).

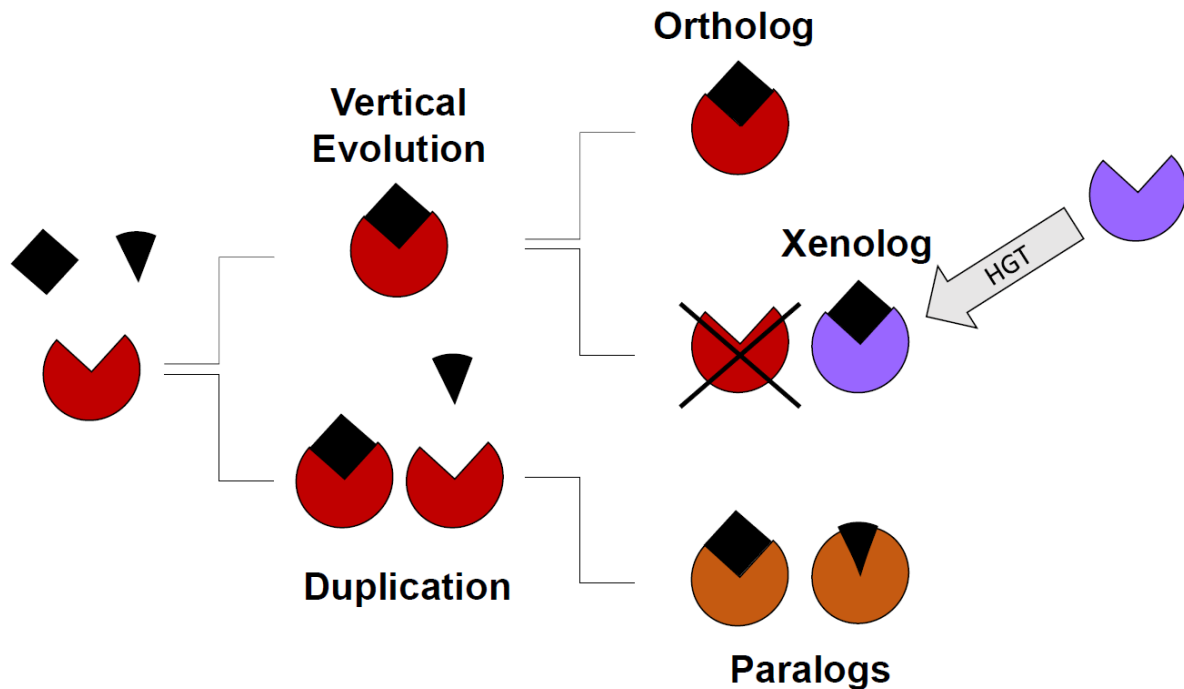


Figure 2. Evolutionary History is Key to Productive Comparative Analysis. This diagram is a basic introduction to evolutionary relationships between homologous proteins that may render them fit or unfit for comparison. The original receptor on the far left can sense one of two ligands. The uppermost path signifies vertical evolution, where the genes encoding this receptor are inherited by the next generation and perform the same function with the same protein. These proteins are **orthologs**, and these are ideal candidates for comparison. However, similar genes can be transferred from other organisms (HGT – Horizontal Gene Transfer, middle right). The horizontally transferred gene may outperform the original in a process termed **xenologous gene displacement**. Comparisons between the original protein and the **xenolog** are inadvisable as separate evolutionary processes have shaped their sequence, structure, and function. The lower pathway signifies a gene duplication event of the original receptor. This scenario is usually transient as evolutionary pressure may be relaxed on both copies, allowing for divergence and gain of a different but related function (*neofunctionalization*). These proteins are now **paralogs**, neither of which would be optimal for comparison to the original receptor.

In comparative genomics, the sources of the sequences in a comparative dataset are as important as the sequences themselves. Bacterial diversity is immense, and the time scales separating the most distantly related branches (phyla) of organisms are vast. Thus, knowing the relationships between the organisms from which the sequences in your dataset originate helps to inform expectations and highlight anomalous findings. For instance, extremely closely related enteric *Proteobacteria* like *E. coli* and *Salmonella enterica* are genetic neighbors, so one would expect their protein sequences to be very similar. Thus, highly divergent regions from related sequences in this case might signal little evolutionary pressure and low functional value. Conversely, comparing an organism from a different phylum (i.e. a deep sea hyperthermophile from *Thermotogae*) to *E. coli*, one would expect extreme levels of divergence, as genes will have had numerous opportunities to adapt to their distinct environments and lifestyles. In this case, highly conserved elements in protein sequences or structures gain elevated importance, as they have stood the test of time.

Understanding how microbes are related to one another is essential for generating hypotheses and interpreting results of all comparative work, and it should be a key factor in determining the composition of comparative datasets. Comparisons can be too recent, as organisms that are closely related have little time to diverge and may not provide statistically significant changes. Any mutations that we observe may just be within the limits of natural variance. Conversely, too deep an analysis, and the comparative targets may no longer have analogous functions or may even have acquired different regulatory or protein-protein interaction partners, creating noise or introducing confounding variables. Because of this issue, one of the first and most important steps in any of our analyses is the generation of a comparative dataset taking into account an appropriate phylogenetic depth. The rapid generation of microbial genomes only stands to make this facet of comparative genomics more informative, as more sequences from diverse groups fill in the existing temporal gaps. These

measures can also help inform evolutionary histories and contribute greatly toward deciphering between the subtypes of homologous relationships discussed in the previous section.

Phylogenetics and Multiple Sequence Alignment. The multiple sequence alignment (MSA) is at the core of the vast majority of comparative work (see **Figure 3** for a simplified example). At the most basic level, it consists of lining up two proteins, one on top of the other, until each given position lines up with the corresponding position on the other protein. When two proteins are used, this can be called a pairwise alignment, and is fairly straightforward to do using a variety of programs or manually. Comparing sequences in this manner is the foundation of phylogenetics, which encompasses all manners of aligning and comparing sequences to determine their relative relationships (or common ancestry). The ease of this process depends on the distance between the two sequences and the percent amino acid identity, and challenging cases may require additional levels of information for clarification (e.g. prior taxonomic relationships as discussed in the previous section).

Automation is required when a large number of sequences are aligned, and the pairwise comparisons become a matrix, where each sequence must be compared against all others to determine the global best fit for all of the sequences. Again, the relatedness can have a dramatic impact on the quality of an alignment. Additionally, the number of sequences can complicate this process or require tremendous computing power. There are many alignment programs, each of which is best suited for different types of alignments. T-COFFEE¹⁰ is one of the best programs for aligning smaller, less conserved sets of sequences, while MUSCLE¹¹ is best suited for large sets of closely related sequences.¹² MAFFT falls somewhere in between, and in our experience is the most consistent alignment program across a wide variety of alignment types.¹³ However, while these assessments are based on subjective observation, objectively, no alignment program is perfect, as no amount of algorithm training can capture every subtle nuance of protein variation that a trained structural biologist can detect.

Conversely, no biologist can manually align large datasets of distantly related sequences, so a tradeoff must be made.

Therefore, it is critical that every alignment that is produced is manually curated. Gaps in alignments are detrimental to the quality of the alignment, but are often present in areas of low conservation such as loop and linker regions. There are many other methods by which to assess the quality of alignments, all of which are grounded in protein structural theory. One commonly accepted method is to look at the hydrophobic and hydrophilic natures of the aligned residues, as both the hydrophobic core and major conserved elements such as beta sheets may serve as recognizable landmarks (**Figure 3**). Mapping secondary structural predictions to an alignment can also be an extremely effective quality control step, and additional programs, like VISSA, are available for this purpose.¹⁴ Finally, the N and C termini of protein sequences can perform critical functions, but in many cases these regions show poor conservation even amongst closely related sequences. Nevertheless, it is important to establish whether or not features such as N-terminal signal peptides or other conserved elements are present before discounting poorly aligning termini. At the N-terminus especially, one of the major unsolved problems in gene calling and annotation is the prediction of start sites, so variability in this region may be unavoidable.

Regardless of the method(s) used to curate an MSA, this method is not only an analytical tool, but also serves as the data input for many deeper and more qualitative downstream analyses (like phylogenetic tree construction or structural visualization). Thus, quality in both the comparative dataset selection and the initial MSA can greatly influence the outcome of an entire project or publication.

Sample Alignment

```
1. ALMTQSFKRIVVMWDELLGGMQKKQRST
2. AILSQTYRKLIVLWDDIMPGMNETSMLI
3. GLMSNSYKNIIVMWD--MPGMQKTQSHV
4. --ITNSYNQMIVVWDELL-GMNRSQ---
5. AL-TQRYKNLIVLWDDIMPGM-ETSLLI
```

```
1. ALMTQSFKRIVVMWDELLGGMQKKQRST
2. AILSQTYRKLIVLWDDIMPGMNETSMLI
3. GLMSNSYKNIIVMWD--MPGMQKTQSHV
4. --ITNSYNQMIVVWDELL-GMNRSQ---
5. AL-TQRYKNLIVLWDDIMPGM-ETSLLI
```

```
1. ALMTQSFKRIVVMWDELLGGMQKKQRST
2. AILSQTYRKLIVLWDDIMPGMNETSMLI
3. GLMSNSYKNIIVMWD--MPGMQKTQSHV
4. --ITNSYNQMIVVWDELL-GMNRSQ---
5. AL-TQRYKNLIVLWDDIMPGM-ETSLLI
```

Figure 3. Sample Multiple Sequence Alignments and Visualization Strategies. This is a sample alignment of five fictitious protein sequences. The top alignment shows a raw, unannotated alignment. Dashes (-) indicate gaps that must be inserted during the alignment process, and can negatively impact the quality of an alignment depending upon their length and their position within the alignment. The middle alignment shows a useful binary color scheme that has coded each amino acid according to their hydrophobicity. Blue residues are generally hydrophobic, whereas red residues are generally hydrophilic. This can be a useful quality control step to manually verify whether or not these properties mesh with the alignment, as the hydrophobicity of residues can determine important structural properties of the overall protein. Finally, the bottom alignment is coded with multiple colors (CLUSTAL scheme),¹⁵ which groups the residues by narrower biochemical properties for finer resolution (i.e. polar residues (green), aromatic residues (orange), charged residues (red +, purple -), hydrophobic (blue), and small turn-like (black)).

Phylogenetic Tree Reconstruction. Phylogenetic tree reconstruction is a quantitative method to analyze and visualize the results of a multiple sequence alignment. Pairwise alignments and evolutionary matrix distance estimations result in the topology of the tree, which is an attempt (and thus an estimation) at reconstructing the true topology.¹⁶ The most widely

used and accepted evolutionary distance matrix is the Jones-Taylor Thomson (JTT) method.¹⁷ The overall topology of the tree and the distances between the branches serve as parameters that a variety of algorithms use to generate trees guided by certain assumptions. In brief, maximum likelihood is a statistical method that selects for the tree most likely to be produced for the given dataset,¹⁸ minimum evolution produces a tree with the fewest number of branch points (e.g. the fewest evolutionary events), and neighbor-joining uses the closest pairs of sequences to iteratively cluster like sequences in branches.¹⁶

Since all methods result in estimations, multiple tree algorithms are commonly used on the same alignment (independently) in order to assess the consistency of their topologies. Furthermore, one can test phylogeny using the bootstrap method, in which the tree is generated n number of times (which can be extremely computationally intensive) to evaluate topological consistency as well. By generating trees 100 or even 1000 times, this method assesses how often branches occur in the same position. Two final considerations for constructing robust phylogenetic trees are the addition of an out-group, and rooting the tree. The former technique allows for distinguishing that a set of sequences are more closely related than a known landmark (often used to localize groups of sequences to certain taxonomic designations). Rooting a tree allows for estimation of a least common ancestor in relation to the sequences, whereas producing an unrooted tree only shows relative relationships within the dataset.

Structural Visualization. Structural visualization is not often paired with sequence analysis, most likely because the vast majority of proteins have not had crystallographic or NMR structures solved. Even when structures are available, proteins are dynamic, so one must realize that any given structure is only a snapshot of a single, stable state. Furthermore, structures are often generated in organisms that are wholly different from where experimental assays were performed. Chemotaxis is a perfect example, where experimental work was performed in *Escherichia coli*, yet crystal structures were obtained using *Thermotoga maritima*

homologs (See **Chapter 2**).^{19,20} Often, model organisms like *E. coli* will have proteins that are too dynamic and are refractory to crystallization, whereas extreme hyperthermophiles like *T. maritima* have evolved to produce thermostable proteins that crystallize far more readily. Unfortunately, as discussed earlier, homologs from phylogenetically distant organisms are not necessarily orthologs and viable comparative targets. In this specific case, the thermostable nature of the protein may have drastically altered the sequence as well.

Additionally, not all structures are of equal quality. Stable protein crystals are difficult to produce and may require significant alterations from native expression, including heterologous expression systems, truncations, modifications, molecular tagging, mutations, and a wide-variety of other experimental techniques required to produce stable crystals. Once crystals are obtained, the resolution of the crystal structure must also be considered, as this determines whether or not one can make conclusions at the amino acid or secondary structure level. Currently, 2 Å resolution is a good cutoff for making residue level observations, while at 3-4 Å resolution, only backbone atoms are reliably placed. Above this level, only macromolecular features may be discernible and sequence visualization is no longer an option.

After all of these considerations, x-ray and NMR crystal structures are invaluable resources. Pairing structures with evolutionary sequence information is a cornerstone of this dissertation work, as it can provide a strategy for relating genomic information to experimental results. At the very least, one can gain an idea of whether or not highly conserved residues play a role in canonical structural roles (i.e. in the hydrophobic core, glycine and proline residues at hairpin turns, etc.), or if these may be mediating protein-protein interactions (i.e. hydrophobic residues that are solvent-exposed). Additionally, in the case of co-crystal structures, one gets a possible snapshot of interaction between two proteins, though whether or not this snapshot makes biological sense may become another matter of debate (more in **Chapter 2**).

Genomic Context. There are two primary methods of using the context of a gene within its genome that can prove useful for comparative analyses. The first of which, **gene neighborhood analysis**, simply involves the gene order, and is a peculiarity of prokaryotes. Often, genes that are part of the same system or interact are coded next to one another or in clusters, which allows for their co-expression. When the number of nucleotides between genes in these clusters is exceedingly small (or even slightly overlapping in some instances), the cluster is referred to as an operon and co-expression and interaction/involvement are almost certainties. Thus, proteins can be “guilty by association”, and by comparing the order of clusters or operons of genes between organisms, one can make systems level analyses. Chemotaxis systems, for example, often occur in operons or clusters, though they may or may not contain chemoreceptors. Conversely, the vast majority of chemoreceptors are found as genomic orphans, in which case gene neighborhood analysis offers little to no insight as to their possible function or chemotaxis system relationships (see **Chapter 4** for more).

The second method, **phylogenetic profiling**, consists of comparing the presence or absence of genes between two or more organisms in order to make phenotype predictions. Most often, pathogens and commensals are compared to identify potential virulence factors or biomarkers that can be used to distinguish them from one another. This type of analysis for single genes is very difficult, as a single gene may not just have one function, but pleiotropic effects within a given organism. Moreover, many diseases are the result of multiple virulence factors and other genes, so the odds of a single gene or even a handful of genes switching the behavior of an organism are unlikely. However, there are interesting trends that have been noticed with chemoreceptor suites, and chemoreceptors are uniquely positioned to heavily influence the environmental niche of an organism. Often, pathogenicity occurs when an organism colonizes a non-native or non-adapted niche, and this type of phenomenon may have contributed to extra-intestinal pathogenicity in *Escherichia coli* (see **Chapter 5** for more).

Materials. As with all experimental methods, raw materials (in this case protein sequences) must be located and collected. Protein sequences are stored and made publicly available by a variety of databases: NCBI (Non-Redundant (NR), RefSeq), UniProtKB (EMBL), and the MiST signal transduction database.²¹ Both RefSeq and UniProtKB are invaluable resources, as they pair experimental data and other valuable information with each sequence. While the scope of the non-redundant database is often useful, the quality and assurance that curated databases provide is often preferable. MiST improves on this by focusing the content on signal transduction and chemotaxis proteins along with providing numerous pre-computed analyses (including domain architecture visualization). As for structures, the Protein Data Bank (PDB) is an invaluable resource as a repository for X-ray crystal and solution NMR generated structures.²² To work using large computational datasets, scripting is an important and necessary ability for any computational biologist or bioinformatician to possess as well. The PERL scripting language was utilized for scripts throughout the dissertation, with each major project mandating several scripts to parse, process, and manage data. While some scripts may be generalized enough to be used on multiple projects (for instance, sequence retrieval using a list of accession numbers or gene identifiers from the NCBI Protein database), most projects require data from different sources using different techniques, requiring new scripts.

Sequence tags are of the utmost importance, and underutilization of these can make follow up analyses more difficult when a protein sequence cannot be readily located or verified. In closed genomes, the most useful type of sequence tag is the locus tag, which provides the genomic location and a species/strain specific identifier to quickly identify where a given sequence came from (e.g. PA2573 is gene #2573 in *Pseudomonas aeruginosa*). Both the accession and GI (GenInfo Identifier) numbers are alternative and widely used identification methodologies, but these numbers change given the status of a genome, and offer no biological information. However, these generic identifiers can be useful for “blinding” the investigator to the

source of a sequence, which can help eliminate biases in certain scenarios. Locus tags are preferable for displaying results and figures though as they are shorter and more useful to the reader, so throughout this dissertation locus tags are used as unique protein identifiers whenever specific gene names are unavailable.

Bioinformatics Analysis and Tools. Unless stated otherwise, there are several typical methods that are used when undertaking comparative genomics work (explanations to follow). Domain architecture and membrane topology predictions were pre-computed in MiST using PFAM 27.0²³ and DAS.²⁴ Multiple sequence alignments were built using MAFFT v7.0 using the I-ins-i algorithm unless otherwise indicated.¹³ Visualization of sequences and pairwise alignments were conducted using JalView v2.8.²⁵ Phylogenetic trees were constructed using MEGA v6.06²⁶ using complete deletion and the JTT substitution matrix.¹⁷ While most bioinformatics tools can be used on a personal computer, computing clusters and parallelization of some tools may be necessary to increase the speed of an analysis or handle large datasets.

As for the software, Jalview is a sequence visualization suite, and is often used as a comparative genomics “workbench”.²⁵ It provides basic pairwise alignment and phylogenetic tree construction functions, in addition to sorting sequences based on a number of helpful properties such as amino acid length. This program allows for modification of sequences and sequence tags, as well as production of alignment figures. Most importantly, it has color coding functions that highlight conservation for given positions and provides consensus and quality scores for given positions. Using the CLUSTAL coloring scheme, residues with like biochemical properties are highlighted, allowing for faster visual recognition of conservation patterns.¹⁵ MEGA is a phylogenetics analysis suite used primarily for phylogenetic tree construction and figure production.²⁶ SeqDepot is an in-house database that contains pre-computed data for protein sequences within the database (Ulrich LE). One unique feature of SeqDepot is the ability of using a protein sequence without any other identifying numbers.

Protein Models. One of the primary tools in our arsenals are Profile and Position-specific scoring matrix (PSSM) Hidden Markov Models (HMMs).²⁷ These are models that are based on multiple sequence alignments of curated remote homologs. In essence, they capture the most conserved positions that relate a given set of proteins for which the model was generated. The hidden Markov aspect takes a given set of inputs and a given set of outputs, and generates an algorithm to approximate the transition from input to output. In this case, the inputs are a series of protein sequences in the form of a multiple sequence alignment, and the outputs are the identity conservation of the amino acids for each given position. In generating the model, the biochemical properties and the identity conservation for a position weight that position more heavily according to conservation of that position as well as the scarcity of residues as well. The relative position of each position is also taken into account, so that gaps may be permitted in certain instances and highly penalized when no gaps occur in the alignment.

These models can be used to search a database of proteins and score each, providing the most related proteins as matches based on scoring cutoffs that are determined specifically for each model (see **Figure 4** for an overview on creating a simple HMM). PFAM is a domain model database where sequences can be queried against models to identify domains. HMMER is a hidden Markov model software package that allows the user to identify domains present in a sequence (HMMsearch) or construct a Profile HMM from an alignment and search against a database with that model (JackHMMER). Hidden Markov Model (HMM) searches in this work were performed using the HMMER3 package.²⁸⁻³⁰

In producing a final score, positions from an alignment with greater conservation are weighted much more highly than non-conserved positions (**Figure 4**). Non-conserved positions

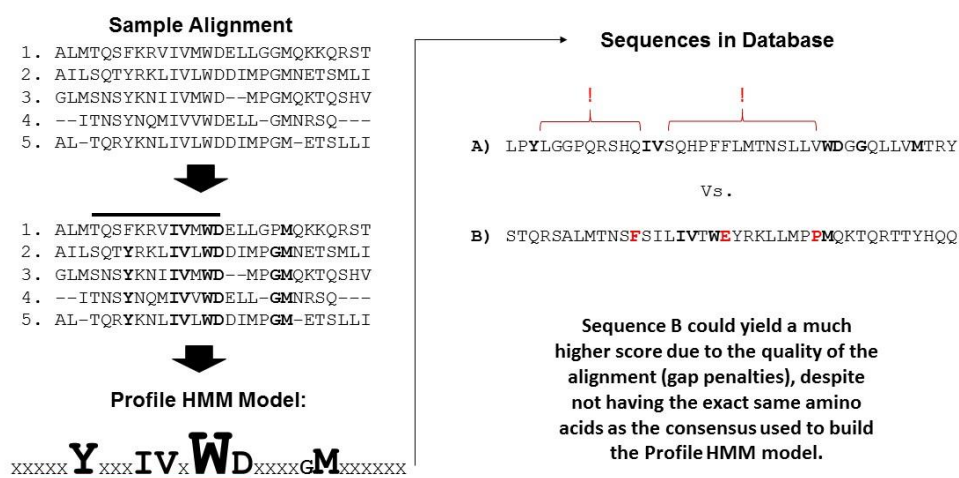


Figure 4. Overview of Producing a Profile HMM and Identifying Matching Sequences. Basic concepts behind generating a Profile HMM and using the model to query a database and identify matching sequences are briefly presented here. From a raw MSA (top left), two major features of the alignment are apparent: residue conservation and gap regions. The middle left alignment highlights these features, with residues conserved at equal to or greater than 80% in bold and a horizontal black bar denoting a stretch of the alignment that is gapless. This information is then used to create the Profile HMM, and a graphical representation is shown in the bottom left (red bar indicates gap penalty region). The model is then used to search against a database of sequences (top right) and report matches.

are simplified as “x” to feature the conserved residues, though the residue distribution of these positions would also be considered in a full model. Additionally, information on the scarcity of these residues is included, as the conservation of a rare amino acid like tryptophan (W) provides more predictive power than a common residue such as aspartic acid (D). The overall summation of scarcity and conservation are represented in **Figure 4** graphically as the size and height of the position, though numerical values would ultimately be used in the model. Finally, gapless regions will factor into the model as scoring penalties imposed on sequences that will

open gaps in that region upon alignment. The final model can then be used to search a sequence database, score each sequence, and report matches based on empirical scoring thresholds that have been calibrated to maximize sensitivity and specificity. Two example sequences are shown in **Figure 4**, where **A** appears to have all of the residues from the model (bold) and **B** has fewer. However, closer inspection shows that **A** has two regions (bracketed with red exclamation marks) that will open up extremely large gaps in the gapless region of the alignment, which would result in significant penalties to the score. Conversely, **B** has biochemically similar residues in the right positions relative to the model, which should result in a substantially higher score than **A**, as well as a potential match if the score exceeds the threshold.

Complementary Experimental Techniques. Several of the key biological events that incorporate chemoreceptors into a chemotaxis complex and allow them to transduce a signal involve protein-protein interaction.³¹ Protein-protein interactions are exceedingly difficult to identify, though a leading method for identifying these regions using sequences is co-evolving residue and compensatory mutation theory.³² Enzymatic and catalytic activity, by comparison, are remarkably easy to recognize from a bioinformatics standpoint, as these processes usually require several adjacent and conserved residues and/or motifs that are highly conserved relative to the structure of the protein. Protein-protein interaction, on the other hand, can be mediated by hydrophobic forces, electrostatic interaction, or van der Waals interactions, none of which are amino acid identity specific. When looking at a sequence, factors like solvent-accessibility and structural context are unavailable, so these regions are variable and difficult to locate. Furthermore, allostereism and cooperativity are not detectable in a sequence, so complexes are exceedingly difficult to characterize if either of these two phenomena play a role in their formation.

Experimental techniques that provide protein-protein interaction information are numerous.³³ The gold-standard is co-crystallization, though this comes with the caveats explained previously. Proteins can also be roughly co-localized using fluorescent protein tags and observed through confocal microscopy, but this is not evidence of interaction alone and more robust methods can be utilized when available. Co-immunoprecipitation is one viable method, where one protein is tagged and affixed to a column to serve as a “hook” to “fish” out interaction partners from an eluted sample. NMR (Nuclear Magnetic Resonance) spectroscopy can be used to observe chemical shifts in functional groups of amino acid residues when two proteins interact in a homogenous solution. Yeast 2-Hybrid capture/prey systems are also widely utilized, but also require both interaction partners to be known beforehand. Each partner is fused to half of a galactose promoter, which in turn controls a lactose operon reporter when it is an intact and functional protein. When the partners interact, the full galactose promoter can induce lac operon transcription of the reporter genes. Finally, FRET (Fluorescence Resonance Energy Transfer) pairs can establish the proximity of two residues within the same protein or between two interacting protein pairs and this technique complemented our results through a companion paper in **Chapter 2**.³⁴ Each residue of a FRET pair is tagged with either a fluorophore or a quenching group. When close enough in proximity, a fluorescent signature change is created by the transfer of the fluorescence energy from fluorophore to quencher.

Comparative genomics is just one aspect of computational biology; there are many computational experimental techniques that provide quantitative means of investigating proteins. Molecular dynamics is one such technique that combines computational power with experimental structural data. Briefly, this method takes molecular coordinates (obtained from x-ray crystal or NMR structures) that represent the atoms of a protein, and then subjects these virtual atoms to equations that describe the laws of motion.³⁵ Additionally, experimentally derived intra and intermolecular forces such as electrostatic interactions and van der Waals

forces are also simulated.³⁶ Computing power is necessary, because all simulated atoms are subjected to “forces” from all other simulated atoms, including simulated water molecules and ions. While there are many experimental parameters that are fairly straightforward to simulate, others require special techniques and considerations (e.g. simulating phospholipid membranes for transmembrane proteins). Ligand and protein docking simulations are useful subdivisions of molecular dynamics, but these require structures with high resolution and minimal synthetic alterations in order to yield biologically relevant simulations. This is a growing field, and as computing power continues to increase, the complexity of the systems that can be simulated increases as well, with cutting edge platforms now limited not by the number of atoms, but by the number of proteins.

As far as ligand recognition is concerned, a co-crystallized ligand within the binding pocket of a protein is a gold standard. Unfortunately, co-crystallizing ligands in their binding pocket can be extremely difficult and time-consuming, and a variety of co-factors, coordinating metal ions, and other conditions may make this an elusive quest. Isothermal titration calorimetry (ITC) and surface plasmon resonance (SPR) are alternative methods that utilize different techniques to detect binding and quantify binding events, and our bioinformatics results complemented ITC/SPR results in **Chapter 3**. Briefly, ITC measures the free energy change when ligand is introduced to substrate, and a change in heat absorbed or released indicates a change in free energy (hence binding).³⁷ SPR, on the other hand involves the affixing of a receptor to a gold plated chip, flowing ligand over the chip, and detecting electrical oscillations (plasmons) triggered by binding events.³⁸

While many of the aforementioned techniques can be extremely informative by themselves, combining targeted mutation studies with these or other techniques is the most complementary experimental strategy available to comparative genomics. Often, our alignments produce conserved residues that we may hypothesize contribute to a certain aspect of that

protein's physiology. When we are able to systematically substitute amino acids with various (and usually opposing biochemical properties), we can validate or disprove our hypotheses (provided that there is an observable phenotype to assay and the mutation does not dramatically alter the "native" structure). For instance, if we predict that a conserved residue in a binding pocket mediates binding, our collaborator can use SPR to confirm binding, mutate that residue to an opposing residue, and assess for binding again.

Chemotaxis, Chemoreceptors, and a House Divided

Before diving into the division facing the current chemoreceptor landscape, one must first appreciate the basics of how chemoreceptors were discovered and how these proteins function within the context of the chemotaxis system in general. Chemotaxis, the ability of bacteria to sense and move toward specific chemical signals, was first recognized and reported in the 1960s by Julius Adler.^{39,40} The discovery of chemoreceptors was instrumental to this breakthrough, as these proteins conferred the specificities necessary for systematically probing the behavior.⁴¹ Using *E. coli* as a model system, five chemoreceptors were eventually discovered and characterized: *tsr*, *tar*, *trg*, *tap* and *aer*. After decades of study, these chemoreceptors have been shown to be dynamic proteins that respond to multiple different signal specificities and can directly or indirectly sense ligands (with indirect mechanisms involving carrier proteins). An entire review could be devoted to just the study of these five proteins, so only very basic details of these receptors will be outlined here.

Tsr and *tar* were characterized by Koshland *et al.* and shown to mediate taxis toward serine and aspartate respectively.⁴² *Tsr* is encoded as a genomic orphan, while *tar* is encoded within the chemotaxis gene clusters, but both have been shown to be major receptors that are indispensable for chemotactic response. *Trg* was shown to mediate taxis toward the sugars

ribose and galactose, which was later shown to occur through indirect ribose and galactose binding proteins respectively.⁴³ *Tap* is a tandem duplication of *tar*, as both genes share tremendous homology and are adjacently encoded within the genome.⁴⁴ This chemoreceptor was later shown to mediate taxis towards dipeptides and pyrimidines.^{45,46} Finally, *aer* encodes an aerotaxis transducer that modulates responses to changes in oxygen concentrations through redox sensing.^{47,48}

As for the breadth of chemoreceptor study, the chemotaxis community has investigated numerous other model organisms with significant and direct impacts on human health and the environment (*Salmonella*, *Rhodobacter*, *Thermotoga*, *Pseudomonas*, *Bacillus subtilis*, *Campylobacter jejuni*, and *Helicobacter pylori* to list a few). Even within these prominent systems, only a handful of other chemoreceptors have been characterized. As of August 2014, there are over 58,000 predicted chemoreceptors in the RefSeq database,⁴⁹ which is a conservative estimate based only on high quality sequences. We have barely scraped the surface across the entirety of chemoreceptor space and there is much more to learn about this major class of proteins.

Fortunately and quite serendipitously though, *E. coli* still serves as a simple and fairly representative chemotaxis system,⁵⁰ so it is still useful to introduce the chemotaxis system as a whole using *E. coli*. There are five core chemotaxis components: chemoreceptors, the adaptor protein CheW, the histidine kinase CheA, the response regulator CheY, and the flagellar motor protein FliM.^{50,51} Traditional chemoreceptors are localized to the periplasmic membrane, where they sense ligands or other stimuli and transduce a signal to the *chemotaxis complex* formed by the highly conserved signaling domain of the chemoreceptor, CheW, and CheA. CheA's kinase activity controls the phosphorylation levels of CheY, and phospho-CheY in turn interacts with FliM at the flagella motor to control its rotation. CheY. Depending on the clockwise or counterclockwise rotation of the flagella, the bacteria will either “run or tumble”, producing a

random walk that can be biased towards attractants and away from repellents. Depending on the type of chemotaxis system involved, numerous phosphatases (CheC/CheZ), methylesterases (CheB), and methyltransferases (CheR) regulate the system downstream from the chemoreceptor:ligand interaction.

Chemoreceptors (MCPs). The methyl-accepting chemotaxis protein (MCP) is a multi-domain homodimer that is identified by its highly conserved signaling tip region (HCD domain or MCPsignal domain). Chemoreceptors are among the most highly variable classes of proteins, allowing for almost any conceivable alteration of domain architecture provided that the signaling domain remains intact. Tremendous evolutionary pressure on this region is believed to be a result of this region modulating multiple structural and functional aspects of the chemoreceptor. Structurally, the sequence of this region produces helical heptad secondary structure, tertiary folding where the signaling tip makes a hairpin turn at 3 conserved glycine residues and the helical heptads form antiparallel stacked helices, and quaternary structure in homodimerization to form a four helix signaling bundle.⁵² This four helix bundle then further complexes with two more chemoreceptor dimers to form a trimers of dimers. Functionally, the receptors must maintain interactions with the adaptor protein CheW and the histidine kinase CheA, and must also transduce signal over several hundred amino acids of length to modulate the phosphorylation activity of CheA (see **Figure 5**).

Chemoreceptors are often referred to as MCPs, as they can be methylated by the methyltransferase CheR and demethylated by the methylesterase CheB. These interactions do not occur in all chemoreceptors, but can be detected by the presence of conserved glutamate residues, which undergo reversible methylation to become glutamate 5 methyl-esters (gamma glutamyl-methyl esters).^{53,54} These modifications of the chemoreceptor result in a remarkable adaptation system that has been referred to as “molecular memory,” allowing the

chemoreceptors to possess a broad dynamic range of sensory capabilities that can tune out high concentrations of one signal to detect exceptionally low concentrations of another.

Other features of the chemoreceptor include the sensory region and the HAMP domain. Sensory regions (LBDs) are numerous and varied in size and architecture, as evidenced by the diversity of commonly detected signal transduction domains. The vast majority of the known sensory specificities will be covered in the section on microbial pathogenicity and various chapters in the main body. The HAMP domain is a ubiquitous helical signal transduction domain that contributes to the propagation of the signal from the membrane to the conserved signaling domain and will not be discussed.⁵⁵

Cytoplasmic Receptors. Not all chemoreceptors are localized to the membrane, as many chemoreceptors clearly lack predicted transmembrane regions. However, very little is known about how these receptors function or how they contribute to chemotaxis as a whole, so their presence in many of the projects in this dissertation has been a source of both challenges as well as new discoveries and leads. Their solubility alone provides them with the liberty to escape the membrane localization and the chemotaxis cluster or array, which creates the possibility for mobile chemotaxis sensors (see **Figure 5**). In fact, soluble chemotaxis clusters have been detected in a handful of organisms including *Rhodobacter sphaeroides*,⁵⁶ but again little is known. One of the best characterized soluble chemoreceptors is AerC from *Azospirillum brasilense*, which dynamically relates the internal metabolic redox state to motility.⁵⁷ Cytoplasmic receptors in *Thermotoga maritima* (**Chapter 2**) and *Pseudomonas aeruginosa* (**Chapter 4**) will be discussed in their respective sections, as well as being prominently featured in **Conclusions (Chapter 6)**.

Histidine Kinase CheA. CheA is a homodimer that can be subdivided into 5 domains (designated by P1, P2, etc.). The P3 domain serves as the dimerization domain. The P5 domain is a structural homolog of CheW, and is the interaction site for the CheA:CheW:Chemoreceptor

complex. P1, P2, and P4 are involved in the autophosphorylation and kinase activity of the protein: the P4 kinase domain phosphorylates a conserved histidine residue on the P1 phosphotransfer domain in an ATP-dependent manner, then the P2 CheY-binding domain stabilizes CheY, where it is then phosphorylated by phospho-P1.³¹ P2 is “suspended” between P1 and P3, the dimerization domain, by flexible, non-conserved linker regions of varying amino acid length. There are many different architectures of CheA (involving the presence or absence of the P2 region, additions of CheY-like response regulator domains, and massive expansions in the phosphotransfer capabilities (CHP system)), allowing for classification of systems based on the CheA protein alone.⁵⁸

The Adaptor Protein, CheW. CheW is (typically) a relatively short (100-150 aa) protein containing a single domain. There are longer homologs with only one CheW domain, and there are also homologs with multiple CheW domains. It has been labeled as a scaffolding adaptor that connects the histidine kinase CheA to the chemoreceptors through protein-protein interaction. It has pseudosymmetry, allowing for interaction with other CheW domains. Since this pseudosymmetry does not limit it to homodimerization, it oligomerizes and forms hexagonal rings.⁵⁹ As mentioned previously, the P5 domain of CheA is homologous to CheW, allowing for incorporation of CheA into the hexagonal ring. As a consequence of this, *in vivo* chemotaxis clusters and *in vitro* complexes show a stoichiometry⁶⁰ of 3 Receptors: 2 CheWs: 1 CheA, which result in each hexagonal unit containing four CheWs and two CheAs with a trimer of dimeric receptors in the middle of the hexagon.^{60,61} A second major class of adaptor, CheV, consists of a CheW domain paired with a response regulator domain, but the full contribution of CheV to chemotaxis has not yet been fully elucidated.

Signal Transduction - Chemoreceptors from a Structural and Mechanistic Viewpoint

In the early years of chemoreceptor study, fundamental mechanisms of how chemotaxis proteins functioned were a major area of investigation. However, DE Koshland Jr. brought major

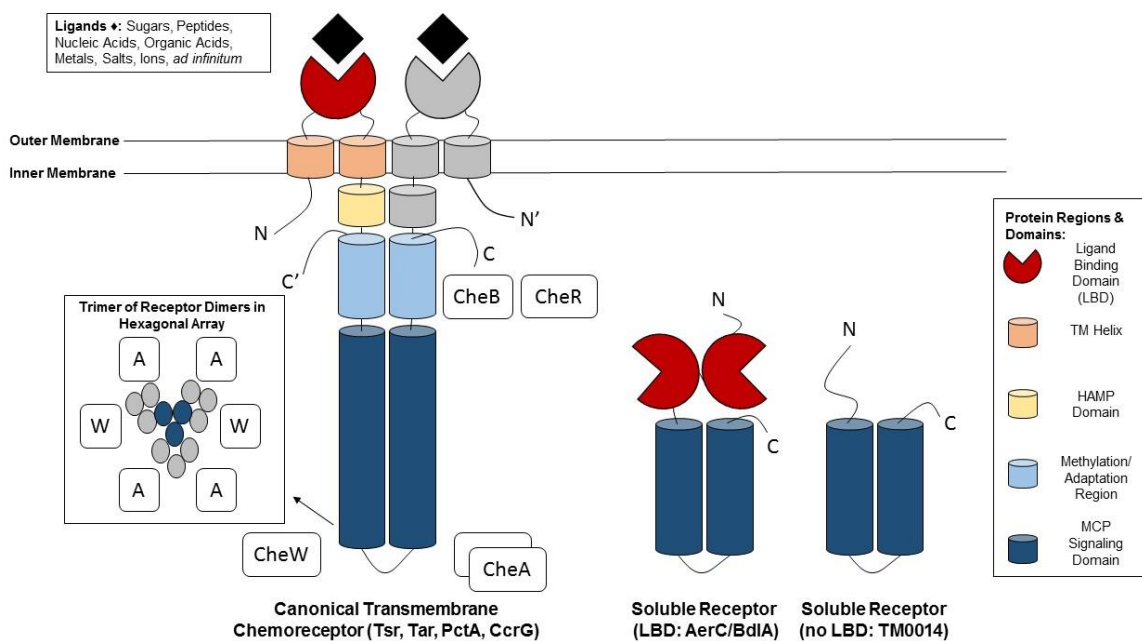


Figure 5. Diagram of Canonical and Non-Canonical Chemoreceptor Localization. The canonical transmembrane receptor homodimer (dimer partner colored in gray) is shown localized to the periplasmic membrane along with its defined domain architecture. The ligand binding domain senses extracellular small ligands in this case. Interacting proteins (adaptor CheW, histidine kinase CheA (a homodimer), methyltransferase CheR, and methylesterase CheB) are simplified and not to scale. The insert is a top down view of the trimer of dimers positioned within a hexagonal array of CheW domains (some of which belong to CheA).⁵⁹⁻⁶² Two soluble chemoreceptor paradigms are also shown (as monomers). Soluble receptors can have defined LBDs, but this is not always the case. Soluble receptors without LBDs may be paired with a partner sensory protein to form a bipartite chemoreceptor.⁶³ Non-canonical receptors from either case are poorly characterized and understood.

awareness to the potential for chemoreceptors to also serve as a model signal transduction system, as they function analogously to eukaryotic neurotransmitter receptors (i.e. G-protein coupled receptors).⁶⁴ One of his major contributions to this type of work was the development of a piston-shift model relating to how the ligand binding domain signals through the transmembrane (demonstrating that the transmembrane alpha helices shift 3-4 Å down, like a

piston).⁶⁵ For many in the chemotaxis field, the mechanisms involved in transmembrane signaling, methylation and adaptation, and protein-protein interactions between the chemoreceptor signaling domain, CheW, and CheA would quickly overshadow how these systems connected to the organism's behavior outside of chemotaxis. This work was quickly confined to *E. coli*, *Salmonella enterica*, *Bacillus subtilis*, and *T. maritima*, four organisms with single chemotaxis systems and relatively small numbers (5-11) chemoreceptors, producing high quality structural and molecular models of chemoreceptor signaling function.

Further Signal Propagation Domains and Models. While the piston movement model served as a starting point, it only focused on transmembrane signaling. In most chemoreceptors, several hundred amino acids must be involved in signaling as well to reach the conserved signaling domain at the receptor tip. JS Parkinson's work on the helical HAMP domain (which directly precedes the transmembrane region), has expanded our understanding of how the signal propagates further down the receptor.⁶⁶ Additionally, the Parkinson lab has been instrumental in performing mutational and cross-linking studies that contributed to modeling the trimer of dimers concept (including highlighting critical conserved positions for this process).⁶⁷⁻⁶⁹ JJ Falke has also been instrumental in investigating the interactions of chemoreceptor signaling, including FRET based and cross-linking studies that have suggested numerous interaction sites.⁷⁰ He also has done major work examining the chemoreceptor structure and theoretically modeling signal propagation as a "yin-yang" mechanism that highlights the tendency of chemoreceptors to conserve knob-in-socket residues involved in helical packing as a potential signal propagation vehicle.⁷¹

Chemotaxis Complex Stoichiometry. A few of GL Hazelbauer's recent contributions to the field include deciphering the global organization of the chemotaxis array, as well as determining the stoichiometry of the smallest functional chemotactic unit. First, after the establishment of the trimer of dimer organization of chemoreceptors, his group reported that a

trimer of dimer unit could drive chemotaxis as effectively as an entire receptor cluster.⁷² Then, by utilizing *tar* trimer of dimers embedded in innovative nanodisc constructs (to approximate membrane localization), the Hazelbauer lab determined that two trimers of dimers and two CheW proteins are required to allow the histidine kinase activity of a single CheA homodimer.⁶⁰

Residue Level Interactions. FW Dahlquist utilizes NMR to probe the chemical shifts of residues involved during chemoreceptor binding to CheA and CheW. DE Ortega (a member of the Zhulin Lab) collaborated with the Baudry lab and the Dahlquist lab, and by using conserved positions in the CheW protein and molecular dynamics simulations, they were able to show that a conserved salt bridge stabilized the interaction between the chemoreceptor and CheA.⁷³ This was significant because CheW's role as an adaptor and scaffold protein had little mechanistic explanation before this work. Recently, the Dahlquist lab has also published two additional NMR studies that illuminated multiple possible contact sites for both the Chemoreceptor:CheW interface and the Chemoreceptor:CheA interface.^{74,75}

Visualization and Chemotaxis Complex Model Development. Visualizing both the subcellular context and the protein-protein interaction events themselves have also been the subject of several productive pursuits. GJ Jensen has employed cryo-electron tomography and microscopy to show both the universal hexagonal structure of the chemotaxis array across multiple phyla of bacteria as well as many other localization studies in *Caulobacter crescentus* and *Vibrio cholerae*.^{61,76} BR Crane has produced numerous x-ray crystal structures of chemoreceptors, as well as several co-crystal structures of chemotaxis proteins from the hyperthermophile, *Thermotoga maritima*.⁷⁷ This includes two crystal structures containing all 3 interacting partners (CheW:CheA:MCP, PDBID: 3UR1, 4JPB).^{61,78} Additionally, cryo-electron tomography and crystallographic structures have been combined by Manson *et al.* and the Jensen and Crane labs to produce full scale models of the chemotaxis array.^{61,79}

Recent Insight into the Signal Transduction Mechanism. Recently, DR Ortega, JS Parkinson and IB Zhulin used molecular dynamics simulations and mutational studies to interrogate the most conserved residue in the signaling domain.⁸⁰ This residue is a phenylalanine that forms a parallel stacking interaction with its mirroring residue (from the other receptor monomer) inside the bottom of the four-helix bundle just before the hairpin tip. Molecular dynamics simulations indicate that this region undergoes a *cis-trans* ring flip that correlates with the on and off states of the methylation and adaptation region. This was probed computationally by mutating conserved glutamate residues to glutamine residues to simulate different methylation states, mirroring an experimental approach conducted by JJ Falke.⁸¹ The phenylalanine residue was then mutated (in the wet lab) to amino acids with different biochemical properties (non-bulky, non-aromatic), where kinase switching activity was no longer observed while expression and structure were unaltered by the mutations. This rotameric switch may drive kinase activity on and off, as it appears to influence how tightly the four helical bundle of the chemoreceptor dimer is clustered. A more robust molecular mechanism may soon result from this and related works including recent findings from Pedetta *et al.*⁸², making this a major discovery in the long history of using chemoreceptors as a model for signal transduction.

Microbial Behavior and Pathogenicity

Both a number of pathogens and non-pathogens have had chemotaxis systems characterized to varying degrees. The paradigm for this work is more concerned with observable phenotypes such as correlating ligand sensing to chemotaxis system output to behavior. However, many pathogens are non-motile, so chemotaxis is often overlooked for other virulence factors such as adhesins, pili, toxins, etc. A common theme of pathogenesis is that of a well-behaved commensal moving from its favored location in the body to a new location where it then proceeds to wreak havoc. Chemotaxis is a likely suspect for influencing the transition of motile organisms from commensal-favoring environments to pathogen-favoring

environments, and it is surprising how poorly understood this connection is in most diseases. Pathogens with an established link between chemotaxis and virulence/pathogenicity are *Pseudomonas aeruginosa*,^{83,84} *Vibrio cholerae*,⁸⁵ *Campylobacter jejuni*,⁸⁶ *Helicobacter pylori*,⁸⁷ *Treponema pallidum*,⁸⁸ and *Borrelia burgdorferi*.⁸⁹ Work from this dissertation in **Chapter 5** also illuminates chemoreceptors' potential contribution to extra-intestinal pathogenicity in *E. coli* (see **Figure 6** for a brief overview of chemotaxis, pathogenic organisms, and sites of infection).

Starting from the left of **Figure 6** and working down, *Borrelia burgdorferi*, the causative agent of Lyme Disease, resides in *Ixodes spp.* ticks and infects humans through tick saliva during prolonged bites.⁸⁹ MNEC (meningitis-associated *Escherichia coli*) infects the meninges (often in neonates),^{88,90} while Spirochete pathogens that cause syphilis (*Treponema pallidum*) and present numerous clinical manifestations including late-stage invasion of meningeal and cardiovascular tissue.⁸⁸ *Pseudomonas aeruginosa* PAO1 is an opportunistic pathogen that complicates respiratory infections, resulting in especially dire prognoses in the young, elderly, immunocompromised, or those suffering from cystic fibrosis. In the stomach, *Helicobacter pylori*, which is a causative agent of gastric ulcers, relies on flagellar motility to colonize the upper region of the stomach, where the environment is amenable to its survival. Additionally, *Vibrio cholerae* and *Campylobacter jejuni* are gastrointestinal pathogens whose chemoreceptors have been linked to contributing to or enhancing pathogenicity. Finally, urinary pathogenic *Escherichia coli* (UPEC) infect the urinary tract by ascending the urethra and ureters and can cause life-threatening infections in severe cases when the kidneys become involved.

Unfortunately, two major factors have complicated this type of research: chemotaxis systems control other systems than just flagella, and there can be multiple chemotaxis systems, both flagellar and non-flagellar. This obscures which receptors talk with which system, as well as what the contribution of these systems is to the overall observed pathogenic behavior.

Chemotaxis Systems Control More Than Just Flagella. Chemotaxis systems have been repurposed in nature to moderate several different behaviors distinct from flagellar motility. The best studied examples of these are type-four pilus mediated twitching motility, alternative cellular function associated with c-di-GMP turnover and biofilm formation, social motility, and fruiting body formation.⁹¹⁻⁹⁴

Type-four pilus motility (TFP or twitching/surface motility), involves the secretion of pili through the periplasm and into extracellular space.⁹⁵ In stark contrast to the rotary propulsion of the flagella, the pilus acts through hydrophobic subunits, which allow for tight adhesion to surfaces. Through retraction of the pilus, the bacteria can essentially drag itself in the direction of the retraction (as the pilus is still attached). The chemotaxis system is involved with modulating pilus synthesis and retraction. The chemoreceptor involved with this system, PilJ, does not have a known ligand specificity though several physicochemical signals (including light) are known to mediate twitching motility.⁹¹

The WSP system from *Pseudomonas aeruginosa* is an example of an alternative cellular function (ACF) chemotaxis system, as its output is not directly related to motility. Harwood *et al.* were the first to characterize this system along with its operon-encoded chemoreceptor, WspA.⁹² This system regulates cyclic-di-GMP turnover. Cyclic-di-GMP is a small molecule unique to bacteria that is used to regulate biofilm formation along with many other signals associated with bacterial virulence.⁹⁶ It does not require the Wsp system for synthesis or control of biofilms, though many organisms have adapted a chemotaxis system to do so.

Multiple Receptors for Multiple Systems. Multiple chemotaxis systems are the exception, not the rule.⁵⁰ However, most of our knowledge of chemotaxis comes from studying the single F7 flagellar system in *E. coli* and *Salmonella*, the single F1 flagellar system in *B. subtilis*, or the

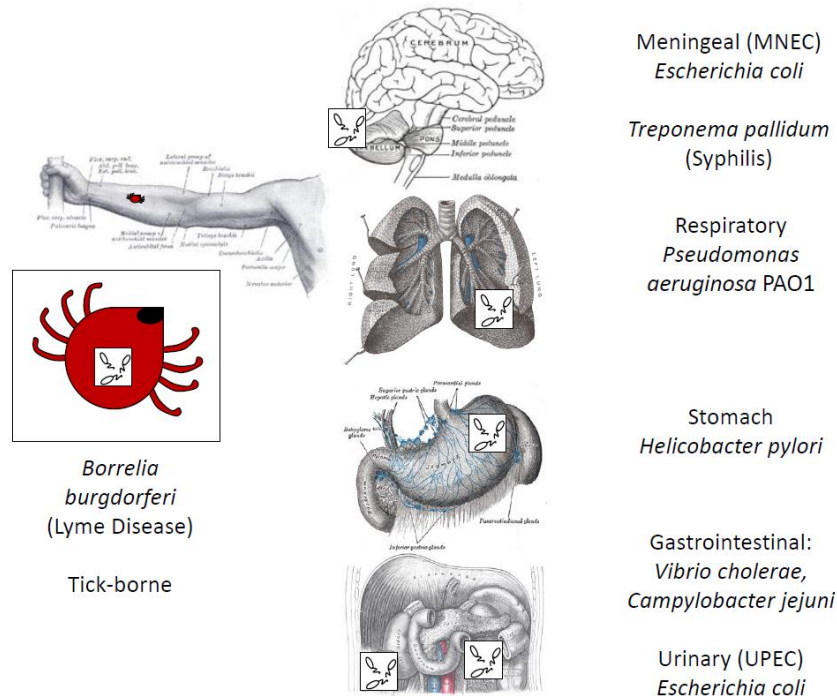


Figure 6. Established Roles of Chemotaxis Systems Contributing to Pathogenicity in Humans.

Multiple strains of motile bacteria have had chemotaxis systems, proteins, or behaviors directly linked to pathogenesis in experimental models of infection. The best characterized examples are illustrated above showing strains and their general site of pathogenicity within the body (anatomical illustrations from Gray's Anatomy).⁹⁷

single F3 flagellar system of *H. pylori* and *C. jejuni*. In all of these cases, one predicted CheA gene implies one chemotaxis system, so there is only one option for chemoreceptor to chemotaxis system connectivity. In cases with multiple chemotaxis systems and a multitude of chemoreceptors, this question becomes much more complicated (**Chapter 4**).

Fortunately, chemotaxis genes tend to cluster together in the genome, and many of the experimentally studied chemotaxis systems in organisms with multiple chemotaxis systems have operons to pair homologs with their cognate system. Even more fortunate, several of these operons contain a receptor (or multiple receptors), which provides a starting point for assigning

pathway connectivity. However, the *vast majority* of chemoreceptors are genomic orphans, and as such have remained uncharacterized. Pathogens such as *V. cholerae* and *P. aeruginosa* contain 45 and 26 chemoreceptors split between 3 and 4 chemotaxis systems respectively, illustrating the problem that faces the chemotaxis community. As it stands, the gold standard for assigning a receptor to a pathway requires enough experimental work to constitute one paper in itself, and at this pace, the majority of chemoreceptors will remain unstudied.

What follows is a comprehensive review of the state of the art in experimentally investigating organisms with multiple chemotaxis systems. Due to the number of studies covered, these are organized by organism, and these sections illustrate both the progress that has been made in assigning receptors to pathways and also the difficulties in doing so with current experimental methods. *Vibrio cholerae*, a gram-negative member of γ -*Proteobacteria* and human pathogen (cholera), serves as an example where little had been reported on the multiple chemotaxis system question until very recently,⁹⁸ because prior *in vitro* work in *Vibrio cholerae* had focused primarily on one chemotaxis system (operon 2).^{99,100} *Pseudomonas aeruginosa* is covered here, but will also be covered in greater detail in **Chapter 4**.

Rhodobacter sphaeroides is a purple, non-sulphur member of α -*Proteobacteria* that grows aerobically, anaerobically, or photoheterotrophically.¹⁰¹ This organism has multiple chemotaxis systems, making it difficult to assign CheA, CheW, and MCP homologues to pathways.^{101,102} McpG in *Rhodobacter sphaeroides* is a membrane-spanning chemoreceptor that localizes to the cell pole and depends on Che proteins encoded by Operon 2, but not homologues encoded by Operon 1.¹⁰³ In order to investigate this, McpG-GFP fusion construct was introduced in place of the McpG gene. Then, a strain lacking the chemotaxis operon was created with the McpG-GFP chemoreceptor. The Operon 1 deletion strain maintained the wild-type localization, while the Operon 2 deletion strain resulted in dispersed localization. Therefore, the authors concluded that

CheW2, CheW3, and CheA2 from Operon 2 are required for proper localization of McpG, successfully assigning one chemoreceptor to one specific chemotaxis system.

For a second chemoreceptor, TlpC in *Rhodobacter sphaeroides* is a soluble receptor that localizes to a cytoplasmic cluster and electronic dense region, not the polar localization shown by McpG. This receptor was shown to depend on CheW3, CheW4, and CheA2 using a slightly different methodology.⁵⁶ In brief, a TlpC-GFP fusion took the place of the original TlpC gene. Then, insertions were created in all of the CheA and CheW homologues. Fluorescence localization intensity and placement were compared to the wild-type for each case, and both CheW3 and CheW4 mutants abolished fluorescence while a CheA2 mutant showed an intermediate fluorescence. In this instance, the authors went a step further by using immunogold electron microscopy to establish that the localization state was truly cytoplasmic and not membrane bound. In each case, the rationale was that the GFP fusion protein showed the true localization. This was done over a variety of growth conditions for robustness.

Chemoreceptor to kinase connectivity is not the only combinatorial problem facing multiple system organisms: there can be multiple CheY homologs as well. In 2009, a mathematical modeling and control systems theory based approach was used to invalidate models of CheA:CheY signaling connectivity. The models were fit to experimental data, and one model was unable to be invalidated. However, chemoreceptors did not factor into this work.¹⁰⁴ In 2010, a computational method using clustering techniques was used to sort and pair chemotaxis clusters and genes (CheABRWY). Chemoreceptors were also not used in this analysis.¹⁰⁵ Thus, assigning chemoreceptors to pathways can be a tremendous endeavor that may not end even once a CheA:CheW:Chemoreceptor association is achieved.

Geobacter spp. are gram-negative members of δ -*Proteobacteria* classified as facultative anaerobes.¹⁰⁶ In *Geobacter*, Weis *et al.* used bioinformatics methods to describe 6-7 chemotaxis clusters in 3 different *Geobacter* spp. They obtained homologues using BLAST and PSI-BLAST¹⁰⁷

against the 3 genomes, using queries from *E. coli*, *B. subtilis*, and *T. maritima*. Several of the clusters contained MCPs, but there were more MCPs outside of chemotaxis clusters than inside. They used multiple-sequence alignments to assign helical heptad classes to chemoreceptor based on the work of Alexander and Zhulin,¹⁰⁸ and postulated that MCPs with like heptad classes/lengths cluster together to diminish unwanted crosstalk.¹⁰⁶

Pseudomonas aeruginosa is a gram-negative γ -*Proteobacteria* that is an opportunistic pathogens of many eukaryotic species including humans.^{91,109} *Pseudomonas aeruginosa* has four chemotaxis systems: Wsp⁹² (ACF), Chp/Pil⁹¹ (TFP), and flagellar systems F6/F7.^{110,111} The F6 system is associated with flagellar motility, and it was linked to a *cheV cheR* cluster.¹¹² The F7 system has been shown to be preferentially expressed during periods of stress, and F7 and F6 systems do not appear to co-localize (CheA-YFP from F6 and CheY2-YFP from F7 do not co-localize).¹¹¹ That said, in a previous paper, CheB2 from the F7 system has been shown to complement CheB1 deletion strains, shows general chemotaxis deficits when CheB2 is deleted, and has been linked to pathogenicity.⁸³ The first two of the previous findings led Harwood *et al.* to posit that the polar cluster is remodeled during periods of stress.¹¹¹

Three paralogous (Cache_1 containing) transmembrane receptors, PctABC, were determined to mediate taxis towards amino acids.¹¹²⁻¹¹⁵ PctABC localize to the cell pole. WspA, conversely localizes to the sides of the organism. In order to probe the localization of WspA and PctABC receptors, the chemoreceptor HCD was swapped between the two types, and their localizations reversed, showing that this broad region was necessary and sufficient to localize each to their wild-type position.¹¹⁶ As such, there seems to be little doubt that PctABC work with F6 and WspA and PilJ work with the ACF and TFP systems respectively. The major factor that has not been researched here is how the F6 and F7 systems share control of the flagella.

Additional Chemoreceptor Work from Pseudomonads. Two aerotaxis chemoreceptors were identified and found to depend on Che Cluster 1 and a *cheR* that is in its own cluster (5) with

cheV. All 5 Che Clusters were assayed with deletion-insertion mutations in order to arrive at a specific system designation.¹¹⁷ A malate receptor (PA2652) in *Pseudomonas aeruginosa* was identified through double knockout mutants, swarm plate assays, capillary tube assays, and complementation. No association with chemotaxis systems was directly assessed in this work.¹¹⁸ Pput_0623 was demonstrated to provide chemotaxis towards cytosine and pyrimidines through capillary assays, complementation, and cross-species knock-in (*Pseudomonas aeruginosa* PAO1).¹¹⁹ Again, no association with chemotaxis systems was experimentally explored (though it stands to reason that it can function with a system shared by the two organisms).

In *Pseudomonas aeruginosa*, a soluble chemoreceptor, BdlA, controls biofilm dispersion. The phenotype was established using 3D structural analysis of the biofilm, and mutants defective in both twitching motility and swimming motility did not appear to affect the dispersion phenotype.¹²⁰ While it makes sense that this chemoreceptor could be involved with the WSP system, to the best of our knowledge it has not been experimentally demonstrated yet that it interacts or localizes with this or any other system. Another soluble receptor, McpS (PA1930), was shown to localize to the polar chemotaxis array (unlike TlpC in *Rhodobacter*). Overexpression caused loss of polar clustering of general MCP population in a dose-dependent negative effect. Cell fractionation confirmed solubility, indicating that another mechanism localizes the receptor to the array aside from the TM regions.¹²¹ Multiple systems were not experimentally investigated in this work as well.

Myxococcus xanthus is a gram-negative soil dwelling δ -*Proteobacteria* with 8 chemotaxis systems and a complex lifestyle, serving as an extreme example.¹²² Through bioinformatics (MSA/Trees), Moine *et al.* phylogenetically clustered this organism's chemoreceptors. They then conducted co-localization experiments to determine if receptors that branched together localized together in the cell.¹²³ Two receptors in *M. xanthus* have been further experimentally characterized (FrzCD and DifA). FrzCD is a cytoplasmic, bipartite chemoreceptor that was

confirmed to participate in the Frz system using in-frame deletion mutants for all Frz genes. Furthermore, the C-terminal region was deleted, which locked the signaling in a constitutive state that abolished movement.⁹⁴ Its dynamic localization profile has been experimentally reported as well.¹²² As for DifA, both a Tn5 deletion, an in-frame deletion, and another separate in-frame deletion in the CheA homolog of the Dif system resulted in the same social motility defects.⁹³

Sinorhizobium meliloti is a gram-negative, nitrogen-fixing symbiont that dwells in soil and is a member of α -*Proteobacteria*.¹²⁴ In *Sinorhizobium meliloti*, there are two chemotaxis systems, eight transmembrane receptors, and one cytoplasmic receptor. These receptors have been deleted and various chemotactic deficits occurred.¹²⁴ Furthermore, all of the receptors have been shown to localize to the pole, except McpS, which is generally distributed throughout the organism and is encoded in the *che2* operon on the symbiotic plasmid pSymA. Three of the transmembrane, polar receptors and the cytoplasmic receptor were co-localized with the major chemotaxis operon's CheA, which the authors took as evidence for this being the cognate system.¹²⁵

Synechocystis PCC6803 is a photosynthetic and phototactic Cyanobacterium.¹²⁶ In *Synechocystis* sp. PCC 6803, mutation of a chromophore containing chemoreceptor (along with mutations in several of the Che genes in an operon with it), all resulted in diminished phototaxis.¹²⁶ These results were confirmed in a slightly later paper that described the chemotaxis system involved in phototaxis to be a type IV pili system.¹²⁷ Cyanobacterial chemoreceptors served as one of the first model systems to explore chemoreceptors in their genomic context within multiple chemotaxis systems, as there are a limited number of diverse chemoreceptors that are encoded within chemotaxis system operons.¹²⁸

Chemoreceptors, Behavior, and Physiology Conclusion. To summarize, there are two pathways currently available to experimentally connect a receptor to a chemotaxis system: 1) if a chemoreceptor is in the operon and there is an observable phenotype, one can conduct deletion studies and then co-localize the proteins. However, this does not rule out cross-talk (especially

with soluble receptors). This approach is also condition-dependent (which many labs take into account, though they could never assay all possible conditions). 2) In a far more likely scenario, if the chemoreceptor is outside of an operon, in addition to doing the above steps, one must include deletion strains of every other CheW, CheA, or Che cluster/operon. The second and much more labor and resource intensive option (evidenced by the work of Armitage *et al.* and Kuroda *et al.* in *Rhodobacter* and *Pseudomonas* respectively) serve as the gold standard for assigning receptors to pathways. It is also important to note that even in the first case, the receptor can be shown to localize to a different system (*P. aeruginosa* McpA PA0180).¹¹¹ Therefore, the experimental cost for assessing orphan receptors is at best a combinatorial exercise, where, for each receptor, each CheW and CheA must be mutated/deleted to establish selectivity. For example, *P. aeruginosa* PAO1 has 20 unassigned receptors, 4 CheAs, 8 adaptor proteins (7 CheWs and 1 CheV), even as one of the best studied organisms on the planet. This problem serves as the impetus for **Chapter 4**. The organisms in **Chapters 2, 3, and 5** all come from single chemotaxis organisms, illustrating in a very tangible manner the skew to which research favors the simpler model systems.

Chapter 2: Refining Chemoreceptor Interactions in *Thermotoga maritima*

Reproduced with permission from: **The 3.2 Å Resolution Structure of a Receptor:CheA:CheW Signaling Complex Defines Overlapping Binding Sites and Key Residue Interactions within Bacterial Chemosensory Arrays.** Xiaoxiao Li, Aaron D. Fleetwood, Camille Bayas, Alexandrine M. Bilwes, Davi R. Ortega, Joseph J. Falke, Igor B. Zhulin, and Brian R. Crane

Biochemistry **2013** 52 (22), 3852-3865.⁷⁸ Copyright 2013 American Chemical Society.

<http://dx.doi.org/10.1021/bi400383e>

Author Contributions:

X Li was the first author and performed crystallographic experiments, analysis, and compiled the manuscript along with C Baya, AM Bilwes, and BR Crane (Crane Lab, Cornell University). AD Fleetwood contributed to analysis of crystallographic results, conceptualization and performance of comparative genomics experiments, and discussion. Specifically, AD Fleetwood conducted phylogenetic and bioinformatics analyses of chemoreceptors from phylum *Thermotogae*, developed a script to prospect co-crystal contact interaction sites from PDB files, built multiple sequence alignments, and contributed to evolutionary history analysis including mapping residues to crystal structures for visualization. DR Ortega performed comparative genomics and bioinformatics analyses of CheW proteins from phylum *Thermotogae*, including recognition of second CheW protein and agreement of contact sites with previous experimental results from CheW. IB Zhulin conducted final multiple sequence alignments and phylogenetic analyses and was responsible for final determination of co-evolving residues. JJ Falke contributed to data analysis and discussion and published a companion paper showing experimental data that independently corroborated co-evolving residue contact sites.⁷⁰

Supplementary figures and tables have been reproduced from this publication and are located in Appendix A.

Abstract

Bacterial chemosensory arrays are composed of extended networks of chemoreceptors (or methyl-accepting chemotaxis proteins, MCPs), the histidine kinase CheA, and the adaptor protein CheW. Models of these arrays have been developed from electron cryotomography, crystal structures of binary and ternary complexes, NMR spectroscopy, mutational data and biochemical studies. A new 3.2 Å resolution crystal structure of a *T. maritima* MCP protein interaction region (PIR) in complex with the CheA kinase-regulatory module (P4-P5) and adaptor protein CheW provides sufficient detail to define residue contacts at the interfaces formed among the three proteins. As in a previous 4.5 Å resolution structure, the paralogs CheA-P5 and CheW interact through conserved hydrophobic surfaces at the ends of their β -barrels to form pseudo six-fold symmetric rings in which the two proteins alternate around the circumference. The interface between P5 subdomain 1 and CheW subdomain 2 was anticipated, whereas the related interface between CheW subdomain 1 and P5 has only been observed in these assemblies. The receptor PIR forms an unexpected structure in that the helical hairpin tip of each subunit has “unzipped” to form a continuous α -helix; four such helices associate into a bundle and the tetramers bridge adjacent rings in the lattice through interactions with both P5 and CheW. P5 and CheW bind a receptor helix in a groove of conserved hydrophobic residues between subdomains 1 and 2. P5 binds the helix N-terminal to what would be the tip region (lower site), whereas CheW binds the same helix near the bundle end (upper site), but with inverted polarity compared to P5. Computational genomics demonstrates that residues in the CheW and P5 recognition surfaces, and at the lower, but not upper, receptor binding site, are highly conserved and that binding partners undergo correlated changes in residue identity when comparing different evolutionary classes of chemotaxis proteins. Evolutionary sequence analyses reveal that two distinct CheW adaptors in Thermotogae utilize the analogous recognition motifs to couple different receptor classes to the same CheA kinase. Key residue positions identified by mutagenesis, chemical modification and biophysical approaches also map to these same interfaces. The CheW receptor interactions are well anticipated by existing data and the companion paper (Piasta et al. (2013); Companion paper)³⁴ demonstrates that the P5 receptor contact as defined in this structure forms in native arrays. Known mutational sites further indicate that structural perturbations about these defined interfaces are involved in the regulation of CheA kinase activity by receptors.

Introduction

Bacterial chemotaxis,¹²⁹ the tendency of bacteria to swim toward attractants and away from repellants, has long served as a model system for understanding transmembrane signaling, motility and cellular behavior.^{31,130,131} Moreover, the underlying sensory pathways of chemotaxis are essential for the infectivity of many prokaryotic pathogens such as *Helicobacter pylori* (gastric ulcers and stomach cancers),¹³²⁻¹³⁵ *Vibrio cholerae* (cholera),¹³⁵⁻¹³⁷ and several types of pathogenic Spirochetes (Lyme diseases, dental disease, syphilis).¹³⁸⁻¹⁴⁰ It has become increasingly apparent that the receptors responsible for binding chemoattractants form extended, ordered structures in the cytoplasmic membranes of cells. These chemosensory arrays are primarily composed of the chemoreceptors themselves, also called methyl-accepting chemotaxis proteins (MCPs), the histidine kinase CheA, and the adaptor protein CheW.¹⁴¹⁻¹⁴⁵ Electron cryotomography (ECT) has revealed a hexagonal arrangement for these receptor clusters^{62,146-150} that is based upon a conserved trimeric assembly of chemoreceptors¹³¹ found in species that range from *Proteobacteria* to thermophilic *Thermotogae*.⁵⁹

Although sensing domains differ among MCPs, the receptors all have a similar construction and are exemplified by the four *E. coli* chemoreceptors, Tar, Tsr, Trg, and Tap.^{131,151-153} Dimeric MCPs span the membrane with four helices (TM1, TM2, TM1', and TM2'), bind ligands through a variable amino-terminal extracellular domain and interact with cellular components through a well-conserved carboxy-terminal cytoplasmic domain (MCP_C). MCP_C is linked to TM2 by a short cytoplasmic HAMP domain that is key to transducing signals across the membrane.^{66,154-156} Each MCP_C subunit folds as two long anti-parallel helices that dimerize into a four-helix bundle.^{52,157,158} The region most distal to the membrane (the tip of the bundle) known also as protein interaction region (PIR) or the kinase control module (KCM) interacts with CheA through CheW. At sites ~140-195 Å away from the receptor tip in the so-called “adaptation region”, specific glutamate residues undergo reversible methylation/demethylation (by CheR

and CheB or/and CheD, respectively) to tune receptor activation of CheA.¹⁵⁹⁻¹⁶² Regulation of the CheB methyl-esterase activity by CheA generates feedback control (known as adaptation).

The histidine kinase CheA complexes with receptors and transduces ligand binding events into initiation of an intracellular phosphorelay that ends by regulating rotation of the flagellar motor.¹⁶³⁻¹⁶⁵ CheA is a dimer with each subunit containing five separate functional units (P1 to P5), strung together as distinct domains over the length of the polypeptide.^{19,166-171} P1 contains the substrate histidine autophosphorylated by the kinase domain (P4). P2 docks CheY for phosphotransfer from P1. The last three domains (P3-P4-P5), comprise dimerization, kinase (ATP binding) and receptor-coupling modules, respectively, and their structures have been determined together for *Thermotoga maritima* CheA (CheA Δ 289).¹⁹

The final core component of the signaling ternary complex is the adaptor protein CheW.^{172,173} CheW has the same tandem SH3-domain-like fold as the CheA P5 regulatory domain and conserves two intertwined 5-stranded β -barrels (designated subdomains 1 and 2).^{19,174} The P3-proximal barrel of P5 (subdomain 1) binds CheW through a pseduosymmetric contact that involves conserved hydrophobic residues on each domain.^{157,174-178} Kinetic and genetic studies suggest that CheW and CheA P5 may compete for binding receptors,^{179,180} in keeping with early observations that CheW is required for kinase activation but not inhibition.^{181,182} The structural similarity of CheW and CheA-P5 would support such an assertion and provide a possible mechanism for switching structural states of the assembly.

We produced a model of the Receptor:CheA:CheW cytoplasmic ternary complex by application of site-specific spin labeling with nitroxides and pulsed-dipolar ESR spectroscopy (PDS) to soluble complexes of MCP_C, CheA and CheW from *T. maritima*.²⁰ Overall, the data revealed that the receptor tip binds CheW but also interacts between the P4 and P5 domains of CheA. The PDS structure is surprisingly asymmetric with the receptor stalk aligning along the

CheA dimerization domain and the P1 substrate and P4 kinase domains projected away from the receptor tips.²⁰

More recently we determined the crystallographic structure of a complex between CheA (P4-P5) :CheW and a truncated MCP_C from the *T. maritima* soluble receptor Tm14 (PDB Code 3UR1⁶¹). This structure was fit to electron density from ECT maps of native receptor arrays in the cytoplasmic membranes of cells.⁶¹ The original crystals diffracted to only 4.5 Å resolution but their high solvent content enabled placement of the proteins and domains. The receptor PIR interacts with the surface of CheW expected from our PDS studies,²⁰ but remarkably, the regulatory domain of CheA and CheW form rings of pseudo-hexagonal symmetry that are consistent with the honeycomb receptor lattice observed by ECT and predicted by the domain arrangements found by PDS.^{20,61} The combined methods describe a complex P6 lattice symmetry for the receptor arrays where networked rings of CheA and CheW associate receptor-trimers-of-dimers into a hexagonal lattice that suspends the kinase domains below CheW-P5 rings. We have modeled and refined crystallographic structures against the ECT data and confirmed that the crystallographic assembly states are consistent with the native arrays in the range of 20-30 Å resolution. NMR studies verify interfaces implied by the extended structure^{74,183} and a very similar model has been subsequently published based on independent ECT data.⁷⁹

Despite these advances, the low resolution of the ternary complex structure prevents a detailed description of the molecular interactions within the chemosensory arrays. Indeed, electron density for side chains could not be resolved in the maps, and although the topology of CheA and CheW allow for a largely unambiguous placing of their secondary structure, the register and rotational orientation of the receptor helices was uncertain.⁶¹ Furthermore, the interaction between the receptor tip and the CheA-P5 domain could only be modeled on the CheW receptor interaction because in the crystals P5 interacts with the receptor at a position

that was only assumed to mimic the native association. Greater structural detail is necessary to not only refine the overall architecture of the arrays, but also understand the mechanism for switching activity states. With the aim of improving the ternary complex crystals we have reengineered the receptor fragments by perturbing helix termini involved in lattice contacts. One of these altered fragments consistently produced crystals that diffracted to better than 3.5 Å resolution. The resulting higher resolution structure maintains the P5:CheW rings of the previous structure, but the receptor itself displays an unusual unzipped conformation in which a tetramer of subunits associates the CheA P5 domains and CheW. Although this unzipped structure is not likely found in the membrane arrays, evolutionary analysis of sequence conservation and mutation patterns suggest that the contacts among the receptor, CheA and CheW displayed by the structure are relevant to the native chemosensory system. For CheA, these assertions have largely been confirmed by a companion report to this study.³⁴

Experimental Procedures. Constructs of *T. maritima* Tm14_s with altered termini (residues 107-192, 107-193, 107-194, 106-191, 106-192) compared those that produced the previous 3UR1 structure (107-191) were PCR cloned into vector pET28a (Novagen) and expressed with an N-terminal Histidine₆ tag in *E. coli* strain BL21 (RIL DE3) (Novagen) after induction with IPTG at 18°C and overnight growth for 21 hours. The Tm14_s fragments were purified first with Ni-NTA chromatography, followed by overnight thrombin digestion, and then size-exclusion chromatography (Superdex 75 Hi-load FPLC column in 50 mM NaCl, 100 mM Tris 7.5, 10% glycerol). *T. maritima* CheW and CheA Δ354 (P4P5 domain, residues 355-671) were expressed and purified as described previously.¹⁵⁷

Cubic shaped crystals (50x50x50 μm³) were grown from a mixture of 520 μM Tm14_s, 457 μM CheA Δ354 and 121 μM CheW after 1 month by vapor diffusion from a 2 μl drop (1:1 mixture of protein and reservoir: 500 μl reservoir of 0.2 M sodium acetate trihydrate, 0.1 M Tris (pH 8.5), 15% w/v polyethylene glycol 4,000). Crystals with a similar shape and size as those

derived from Tm14_s (107-191) were grown after 1 month. Among the new crystals (residues 107-192) consistently diffracted to 3.5 Å resolution. Crystals were soaked briefly in cryoprotectant that consisted of 85/15% (v/v) reservoir solution with glycerol prior to data collection in a N₂ cold stream. Diffraction data were collected at 100K with synchrotron radiation at beamline A1 at the Cornell High Energy Synchrotron Source (CHESS). Selenomethionine was also incorporated into the Tm14_s (107-192) to aid in efforts to determine the helical registry, but unfortunately, the selenomethionine incorporated protein did not produce crystals.

Crystal Structure determination and refinement. Diffraction data were processed with HKL2000.¹⁸⁴ Initial phases were obtained with molecular replacement in PHASER¹⁸⁵ employing the 4.5 Å ternary complex structure with Tm14_s (107-191) (PDB 3UR1) as a search model. The Tm14_s subunits were manually unfolded and built into the resulting electron density with XFIT.¹⁸⁶ The resulting structure was refined with PHENIX.¹⁸⁷ B-factor sharpening¹⁸⁸ and composite omit-map calculation in PHENIX approved maps and allowed for proper interpretation of the Tm14_s unzipping.

Bioinformatics software and data sources. Sequences of CheA, CheW and MCP proteins were retrieved from the MiST2 database²¹ using sets of domain definitions that are specific for each protein, as described previously.¹² MCP sequences were classified using custom hidden Markov models.¹⁵³ Pairwise sequence alignments were built using BLAST v.2.2.17¹⁸⁹ with default parameters. Multiple sequence alignments were constructed in MAFFT v.6.0¹⁹⁰ with its L-INS-I algorithm. The conservation pattern was analyzed in Jalview²⁵ using underlying tools. Minimum evolution and maximum-likelihood phylogenetic trees of CheW protein sequences were constructed from the corresponding multiple sequence alignment and analyzed using the MEGA5 package.¹⁹¹ Operons were predicted based on inter-genic distances.¹⁹² A protein cutoff scanning technique¹⁹³ with a beta carbon distance of 6 Å was used to prospect contact sites from PDB files of the co-crystal structures. Measurements were carried out using a custom Perl

script, which compares every other beta carbon's coordinates to the query atom in all-against-all matrix analysis. The distance is computed according to the following equation: Distance = $\text{SQRT}((x_1 - x_2)^2 + (y_1 - y_2)^2 + (z_1 - z_2)^2)$; where x, y, and z are their respective orthogonal coordinates.

Results

Crystal lattice engineering. In efforts to improve the diffraction resolution of the original ternary complex crystals (PDB code 3UR1), modifications were made to the termini of the shortened Tm14 receptor (Tm14_s). The 4.5 Å resolution structure indicated that these termini contacted each other on the symmetry axes of the crystal lattice, and thereby allowed the receptor dimers to stack “end-to-end” with aligned helices (**Figures 7 and 8**). Several new constructs were generated with shifted termini to perturb this principal lattice interaction during crystallization (see Experimental Procedures). One of these (residues 107-192) produced crystals that consistently diffracted to ~3.5 Å resolution (**Table 1**). Ultimately, a new 3.2 Å resolution structure was determined from these crystals by molecular replacement (MR) with the previous 4.5 Å structure as a probe (**Table 1**). MR revealed that the general placement of the CheW, P5 and Tm14_s units were quite similar in the two crystals. Indeed, refinement of the original model against the new data gave $R_{\text{factor}}/R_{\text{free}}$ values of 0.231/0.272 to 3.5 Å resolution. However, examination of the higher resolution electron density revealed that the helix register in the MR model was not compatible with the side-chain electron density, particularly in the bundle core, where two invariant Phe154 residues at the receptor tip could not be placed without offsetting their side chains relative to their position in the mostly complete structure of Tm14 (PDB Code 3G67). This consideration led to a new interpretation of the receptor structure, where the subunits had become unzipped and then associated to form a tetramer of antiparallel helices (**Figure 7**). Refinement of the new configuration gave significantly improved refinement

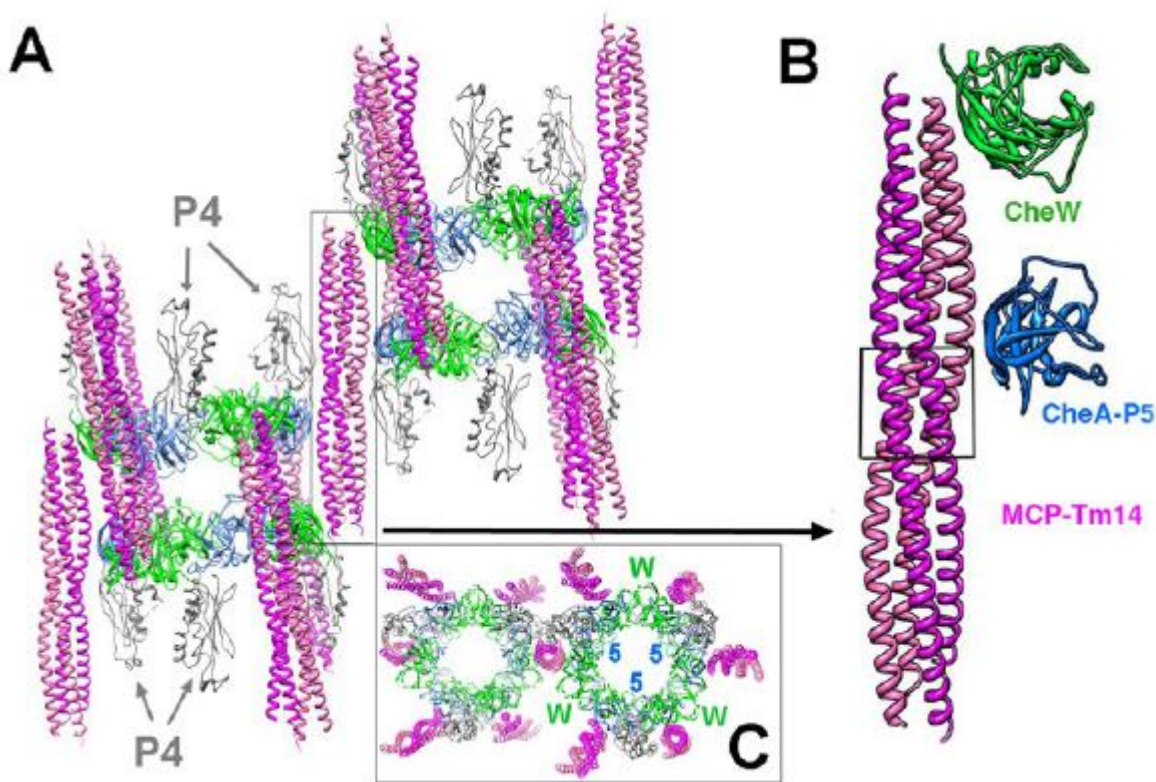


Figure 7. Molecular Composition of the 3.2 Å Resolution Ternary Complex Crystals. (A) Molecular interactions generated by the R32 crystal symmetry. CheW (green) and the CheA P5 domain (blue) form rings of three fold symmetry, with the two proteins alternating around the circumference. The P5-CheW rings are held together by extended 4-helix bundles formed by the Tm14_s receptor (pink and purple). Four Tm14_s subunits have unzipped into continuous helices and associated into a tetrameric 4 helix bundle; each bundle binds two P5 domains near its center (lower site), and two CheW domains at its periphery (upper site). If the pink helices were joined at the middle of the bundle, and the purple helices were similarly joined, they would produce end-to-end MCP hairpin tips as reported in the previous 4.5 Å resolution structure (see **Figure 8**). Tm14_s tetramers interact with every P5 or CheW domain around the ring, although the complete tetramer is only shown for the boxed receptor. The CheA P4 domains (grey) project in large solvent channels above and below the rings although their electron density is not well defined. (B) Close-up of the Tm14_s tetramer and its interaction with one P5 domain near its center and one CheW at its end. The region that would normally form a helical hairpin instead folds as a continuous helix (boxed). (C) View down the central receptor bundle as designated in (A). Two of the four rings associated by the Tm14_s tetramer are shown.

statistics ($R_{\text{factor}}/R_{\text{free}} = 0.203/0.225$; **Table 1**). The unzipped tetramer is surprising because Tm14_s is largely dimeric on purification, as judged by multi-angle light scattering coupled to size-exclusion chromatography (**Figure S1** in Appendix A). In addition, related constructs of Tm14_s have been studied in complex with CheA and CheW by solution NMR, where they are also dimeric.^{74,194} One possible explanation for the unzipping may be linked to construct design. In addition to the native residues of the receptor, the expression vector introduced four non-native residues at the N-termini (Gly-Ser-His-Met-Ser₁₀₇-). Assuming that Ser107 holds the same position in the helical heptad repeat as it does in the structure of the mostly complete Tm14 (3G67), the non-native His would reside in a “d” position, internal to the hydrophobic core of the bundle. Two His side chains (one from each subunit) directed at each other from across from the bundle core would clash, and perhaps their introduction destabilized the Tm14_s dimer under the crystallization conditions.

Molecular arrangements within the ternary complex crystal. Despite the switch from a dimeric to a tetrameric Tm14_s, the arrangement of components in this new structure and 3UR1 are quite similar. The conformational changes in the tip that allow for the unzipping of the helical hairpin and the switch from two end-end dimeric hairpins (as would be found in a typical MCP) to an antiparallel tetramer of unzipped helices are centered in Gly148, and to lesser extent Gly151 (**Figure 8**). The Gly148 ϕ/ψ angles change from an otherwise disallowed region of Ramachandran space, to the helical region; Gly151 changes conformation more slightly to allow for an i to i+4 main-chain hydrogen bond between Ala150 and Phe154. The Gly rearrangements in the extended tip (residues 147-153) produce a typical heptad repeat of 4-helix coiled-coils with Ala147 and Phe154 residing in the most buried “d” position (**Figure 8**).

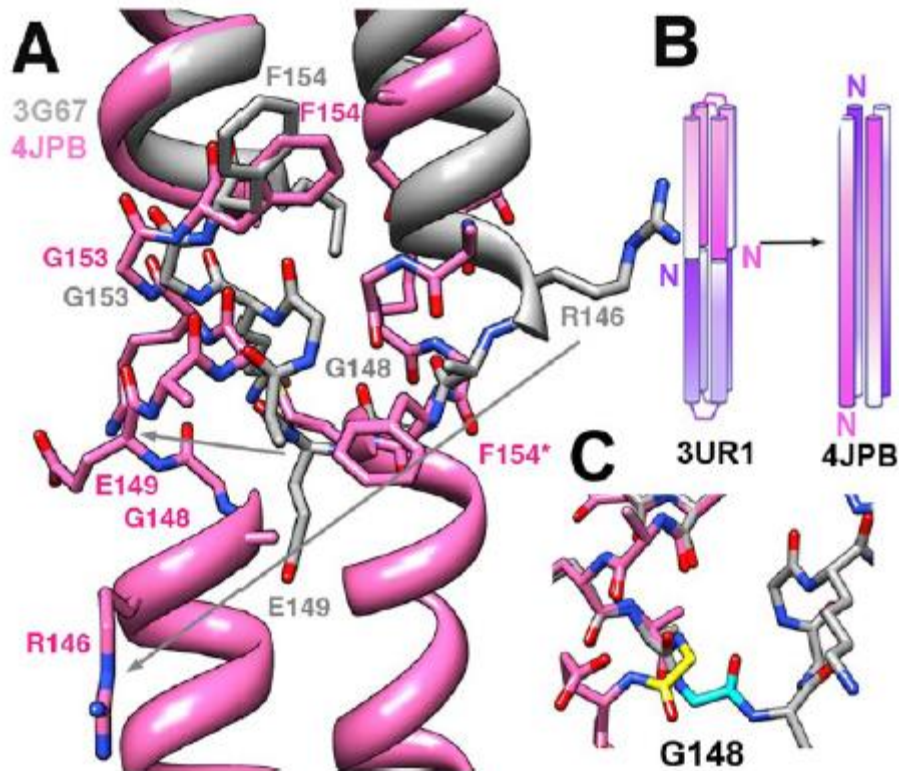


Figure 8. Unzipping of the Tm14_s Helical Hairpin. (A) Superposition of one subunit of the Tm14 structure (3G67, grey) and two subunits of the unzipped Tm14_s (pink). Most of the residues in the transition region maintain similar backbone conformations in both structures, with the exception of Gly 148, whose ϕ/ψ angles change from values disallowed for C β -containing residues to α -helical. Grey arrows map residues in the hairpin turn conformation to the extended conformation. The two Tm14_s helices are antiparallel and offset from one another by two helical turns (note the position of Phe154 and its symmetry mate Phe154* in the opposing subunit). (B) Schematic depicting the relationship between two end-to-end hairpins to a tetrameric 4-helix bundle. Color saturation of the helices decreases from N- to C- termini.

Adjacent antiparallel helices of the tetramer pack similarly as in Tm14 or other MCP structures (3G67, 2CH7, 1QU7, 3ZX6) but the symmetry-related helices from the other two subunits are shifted by approximately one helical turn relative to the first pair. This produces a ladder of the four Phe154 residues from each subunit at the center of the bundle. Importantly, the structure presents a pair of antiparallel helices to CheW and P5 as would be found in other MCP receptor

structures, but the two sides of the bundle (colored purple and pink in **Figs. 7-11**), which can be considered as independent binding surfaces, are offset from one another.

The crystallographic asymmetric unit contains one subunit of CheW and one subunit of the CheA P5 domain, each interacting at a different position on the tetrameric receptor (**Figure 7B**), with P5 close to the center (lower position; **Figure 7B**), and CheW at the end (upper position; **Figure 7B**). Like in the low-resolution structure, the R32 crystal symmetry generates rings of alternating CheW and P5 domains (**Figure 7AC**). Each ring contains 3 copies of CheW and 3 copies of P5 (**Figure 7C**). The paralogs interact through the ends of their β -sheets, with subdomain 1 of P5 binding subdomain 2 of CheW, as previously characterized in complexes of CheA with CheW.^{61,157,174} However, to complete the ring, CheW subdomain 1 also interacts with P5 subdomain 2 in a contact pseudosymmetric to the first (**Figure 9**). Given that CheW and P5 are themselves paralogs, the rings have pseudo six-fold symmetry. A receptor helix associates with the groove between the two β -barrels of each domain and thereby produces pseudo-six-fold symmetric arrangement of receptor bundles. ECT in concert with the low resolution structure suggested that in native membrane arrays, one trimer-of-receptor-dimers associates at each P5 or CheW binding site.^{61,79} Due to membrane incorporation of the native receptors, all of the tips would engage the CheW/P5 rings from the same direction, rather than with the alternating polarity found in these crystal structures. In the unzipped receptor configuration, the Tm14_S helix that primarily binds CheA-P5 does so in a region that would normally be N-terminal to the hairpin tip (this binding helix will henceforth be referred to as the “N-terminal helix”). CheW binds the N-terminal end of this same helix, but does so with the binding groove flipped over relative to that of P5 (**Figure 10**). Thus, in the extended crystal lattice, two CheW/P5 rings, related by twofold rotation, are bound at their edges by six receptor tetramers that alternate their orientation around the ring from “up” to “down” (**Figure 7**).

Table 1. Data Collection and Refinement Statistics for Ternary Complex Crystal Structure

wavelength (Å)	0.97700
spacegroup	R32
cell parameters	a= 213.99 b=213.99 c=208.19
resolution (Å)	46.2–3.2 (3.3–3.2) ^a
no. of observations	169251
no. of unique reflections	30554
redundancy	5.5 (3.9) ^a
completeness (%)	99.7 (99.2) ^a
R _{merge} ^b	0.105 (0.513) ^a
I/σ(I)	20.3 (1.4) ^a
Refinement statistics	
resolution range (Å)	46.2–3.2 (3.3–3.20) ^a
R factor, %	19.6 (33.6) ^a
R _{free} , %	22.0 (36.9) ^a
molecules/ asym unit	1 P4-P5, 1 CheW, 2 Tm14 _S (107-192)
residues/ asym unit	576
solvent content (%)	84
overall B-value (Å ²)	36.1
main-chain B-value (Å ²)	32.9
side-chain B-value (Å ²)	39.5
Wilson B-value (Å ²)	40.5
Geometry	
bonds rmsd (Å)	0.01
angles rmsd (°)	1.33
Ramachandran plot, %	
most favored	89.6
additionally allowed	9.5
generously allowed	0.7
disallowed	0.2

^a Highest resolution range for compiling statistics. ^bR_{merge} = $\sum_i |I_i - \langle I \rangle| / \sum_i I_i$. ^cP5 residue 550 and CheW residue 87 are in disallowed φ/ψ regions but have well-defined electron density.

CheW interaction with CheA-P5. The homology between and pseudosymmetry within P5 and CheW produce rings composed of twelve β -barrels, each β -barrel representing a subdomain from either P5 or CheW (**Figure 8**). At the interfaces formed between subdomains of opposing proteins, three anti-parallel β -strands ($\beta 3'$ - $\beta 4'$ - $\beta 5'$ for subdomain 1, and $\beta 3$ - $\beta 4$ - $\beta 5$ for subdomain 2) wrap conserved, hydrophobic surfaces against each other in an anti-parallel fashion (i.e. the $\beta 3'$ - $\beta 4'$ loop of P5 subdomain 1 interacts with the $\beta 4$ - $\beta 5$ loop of CheW subdomain 2 and $\beta 3'$ - $\beta 4'$ loop of CheW subdomain 1 interacts with the $\beta 4$ - $\beta 5$ loop of P5 subdomain 2; **Figure 9**) Although the Val, Leu, and Ile residues projecting from $\beta 3(\prime)$ - $\beta 4(\prime)$ on all four unique surfaces are quite conserved, the loops connecting the strands differ considerably between the two proteins and also between the two subdomains (**Figure 9**). The sequence and structural variation within these loops likely gives rise to specificity for ring assembly. In agreement with experiments,^{157,174,195} the interface between P5 subdomain 1 and CheW subdomain 2 is predicted to be stronger (880 Å² buried surface area per subunit; ΔG of formation = -13.6 kcal/mol; Specificity P= 0.024¹⁹⁶) compared to that between CheW subdomain 1 and P5 subdomain 2 (591 Å² buried surface area per subunit; ΔG of formation = -4.3 kcal/mol; Specificity P= 0.282). The latter, “weaker” interaction has not been observed outside of the crystallized ternary complexes, although some mutational and modification data suggests that positions on this surface do have a functional role (*vide infra*).¹⁷⁷

MCP interactions with P5 and CheW. The junction between the two subdomains of either P5 or CheW harbor conserved, branched hydrophobic residues (Kinase P5/CheW: kI560/wV27, kI563/wI30, kL457/wL14, kI566/wV33, kL623/wV98) that form a groove to bind the receptor N-terminal helix (**Figure 10AB**). The receptor helix binds into this region on P5 with a row of exposed hydrophobic residues (rIle135, rLeu138, rIle142) as well as rAsn139, which provides two key hydrogen bonds to the peptide backbone of kIle566 on the extension of $\beta 2$ that connects the subdomains (417 Å² buried surface area per subunit; ΔG of formation = -5.2

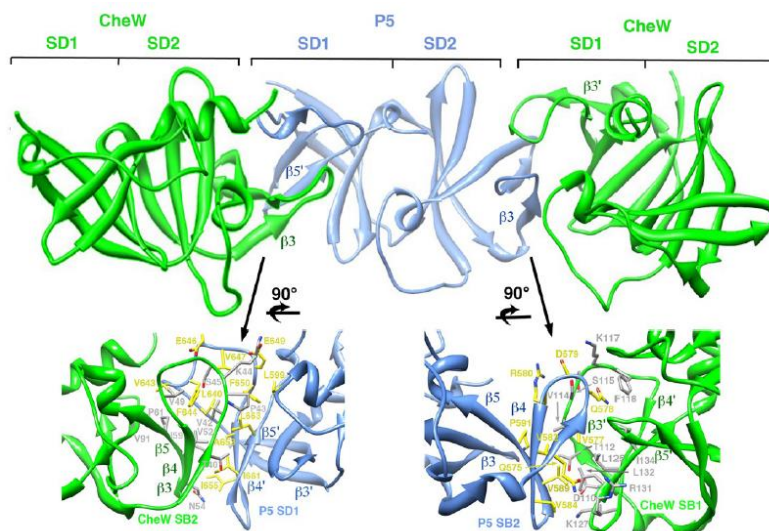


Figure 9. Pseudosymmetric Contacts Made by the CheW and P5 Subdomains. (Top) Half of one P5-CheW ring viewed from the center. (Bottom) The interfaces formed between CheW-subdomain 2 (SB2) and P5-subdomain 1 (SB1) or P5-SB2 and CheW SB1 rotated 90° relative to their orientation above. The two contacts are very similar, both involving close associations of the $\beta 3$ - $\beta 4$ - $\beta 5$ to $\beta 3'$ - $\beta 4'$ - $\beta 5'$ strands on the respective domains and the conserved hydrophobic residues conserved therein. Nonetheless, substantial differences in the $\beta 2(')$ - $\beta 3(')$ loops produce specificity for the interactions.

kcal/mol; $P = 0.287$). Mutagenesis studies have strongly implicated Asn139 in chemoreceptor array structure and function⁶⁷ and the hydrogen bonding interactions it makes likely serve as an anchor for P5 relative to the receptor tip (**Figure 10A**). Another potential anchoring contact involves the side-chain to main-chain hydrogen bonds between rArg146 and the kAsp546 at the periphery of the interface. Mutants of the corresponding Arg residue in Tsr (residue 366) were found to impair or abrogate chemotaxis responses in *E. coli*.¹⁹⁷

The interface between CheW and Tm14_s involves the analogous residues on CheW as on P5 (**Figure 10B**), but the receptor contact forms from a stretch of hydrophobic side chains (rIle109, rLeu113, rIle116) three heptads N-terminal to the P5 contact (331 Å² buried surface area per subunit; ΔG of formation = -4.7 kcal/mol; $P = 0.247$). Moreover, Glu114, which follows the central hydrophobic residue also makes side-chain to main-chain hydrogen bonds

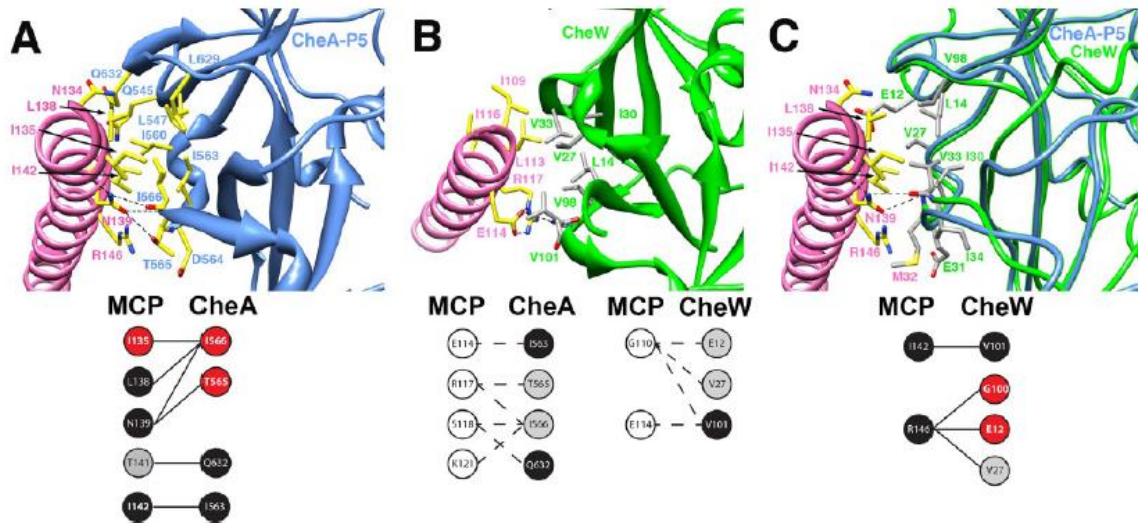


Figure 10. Interactions Between Tm14_s and P5 or CheW. (A) Contact between the N-terminal helix of Tm14_s (magenta, N-terminus up) and the groove between subdomain 1 and 2 of CheA P5 (blue). Residues in the interface (yellow bonds) are primarily hydrophobic, with the exception of Tm14_s rAsn139 which hydrogen bonds with the main-chain of P5 kIle566 on the connection between $\beta 2$ and $\beta 3$ and the kThr565 side chain. Binding spots on MCP for CheA predicted by evolutionary information are given below. White circles represent low sequence conservation in a given position (<80% consensus, no functional conservation); grey circles represent strong sequence conservation (>95% consensus, functional conservation); black circles represent the strongest sequence conservation (100% consensus, identical residues or the same charge conservation); red circles represent positions with correlated mutations. Solid lines identify contacts whose evolutionary history is consistent with the correlated mutation hypothesis; dashed lines identify contacts whose evolutionary history is inconsistent with the correlated mutation hypothesis. (B) Contact between the N-terminal helix of Tm14_s and CheW. The Tm14_s helix (pink) runs in the same direction as in (A), but CheW (green) is rotated $\sim 180^\circ$ relative to P5. Due to the pseudosymmetry of the P5/CheW domains the contact is very similar as in (A), with hydrophobic packing central to the interface and a side-chain-to-main-chain hydrogen bonds between rAsn114, and wVal98, which resides on the connection between $\beta 2'$ and $\beta 3'$. Evolutionary analysis below suggests that this structure does not represent a conserved interaction. (C) Superposition of CheW on to P5 demonstrates that CheW conserves chemical character of the residues (grey) at many of the key positions that mediate the contact between Tm14_s and P5.

with the extension of $\beta 2'$, in close analogy to Asn139 with $\beta 2$ in the P5 contact. Thus, the upper CheW binding surface of Tm14_s has a similar chemical character compared to the lower surface

that engages P5, but the receptor helix runs across the binding groove in the opposite direction. Nonetheless, the intrinsic pseudo-twofold symmetry of CheW that relates the subdomains essentially compensates for the reversed helix and produces an interaction that is quite similar to that seen with P5 at the lower position.

The upper binding site of CheW, and hence the second P5/CheW ring is not accounted for by the current models of array structure, which indicate all of the CheW:P5 units are found in one plane, at the tip of the receptors.^{61,79} It is then likely that the upper CheW binding site facilitated lattice formation by mimicking the natural site and is actually located at the lower position as with P5 (residues r135-r146). In support of this notion, rIle135, rIle136, rLeu138, and rIle142 all undergo chemical shifts in solution NMR studies of CheW binding,⁷⁴ PDS measurements of spin-labeled proteins localize CheW to the lower site²⁰ and genetic and biochemical experiments are consistent with this docking arrangement.^{176,198-200} Indeed, superposition of CheW onto P5 indicates an excellent fit with the Tm14_S 135-146 motif into the CheW groove (**Figure 10C**). However, the strong similarity between the upper and lower receptor binding motifs should not be completely dismissed and raises the possibility of multi-layered rings in other contexts, perhaps involving receptor systems that are not membrane associated.

Computational genomics. We sought to apply a computational genomics approach to independently predict CheW and P5 binding sites on MCP and project the relevance of the interactions found in the ternary complex structure to the greater genomic landscape. Rapid accumulation of genomic data in recent years allows productive comparative sequence analyses to identify evolutionary conserved residues that are important for structure and function including protein-protein interactions sites. A “correlated mutation” hypothesis states that destabilizing changes in one position can be evolutionary fixed by a compensatory modification nearby.³² The relationship between correlated mutations can be derived from

multiple sequence alignments of protein sequences; however, due to the complexity of protein-protein interactions (e.g. mutually dependent residues may not necessarily be in direct contact) there is no single method or approach to successfully predict contact sites. In the case of the chemotaxis system, interacting proteins (MCPs, CheA and CheW) evolve in various subclasses with different protein interaction networks.⁵⁰ This essentially prohibits applying statistical methods, such as “direct coupling analysis”²⁰¹ that rely on very large datasets of uniformly interacting proteins. To circumvent this problem and provide direct correlations to the available structural data, we have carried out comparative genomic analysis of the chemotaxis system of *T. maritima* within the well-defined limits of its specific subclass F1⁵⁰ and taxonomic position (phylum *Thermotogae*).

First, we retrieved sets of CheA, CheW, and MCP protein sequences from all available genomes of organisms from *Thermotogae*. A comprehensive list of these proteins can be found in **Table S1** of Appendix A. Satisfactorily, all genomes of *Thermotogae* contained a single CheA protein that was confidently assigned to the F1 class.

Two distinct CheW proteins are present in Thermotogae genomes. We have analyzed the sets of CheW and MCP sequences to reveal potential diversification within these protein families in *Thermotogae*. Phylogenetic trees constructed from a multiple sequence alignment of CheW protein sequences revealed two distinct sets of orthologs exemplified by *T. maritima* TM0701 (termed CheW1) and TM0718 (termed CheW2): The longest branches on both minimum evolution and maximum likelihood trees separate the two classes (**Figure 11**). This classification is independently validated by the fact that all *cheW1* genes were found in operons together with *cheA* genes, whereas none of the *cheW2* genes were a part of these operons. Strong conservation of sequence and structure of the CheW2 protein suggests that although it has not been successfully experimentally characterized, this protein is functional in *T. maritima* and all its relatives.

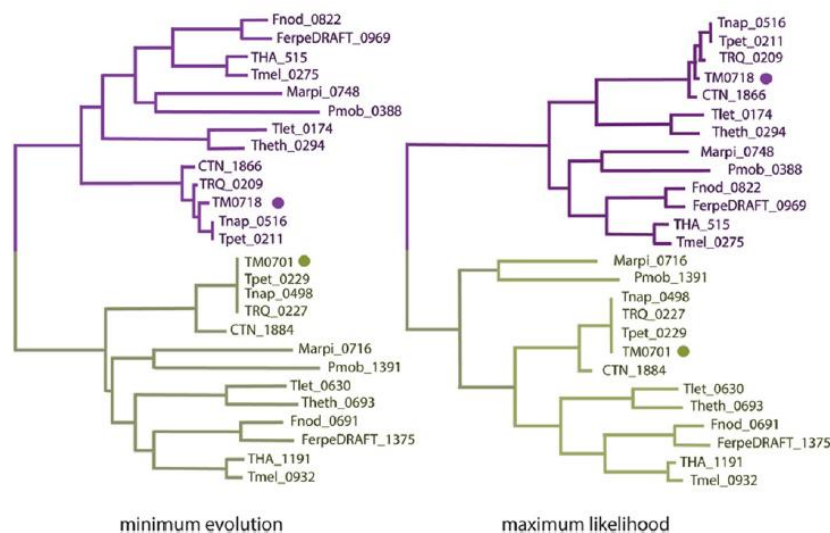


Figure 11. (Minimum Evolution and Maximum Likelihood) Phylogenetic Trees Showing Two Groups of CheW Orthologs from *Thermotogae*. CheW1 group is shown in green and CheW2 group is shown in purple. Sequences are represented by their locus tag numbers. Sequences from *T. maritima* are labeled by circles.

Two distinct MCP classes are present in *Thermotogae* genomes. All 116 MCP sequences from *Thermotogae* were matched against hidden Markov models (HMMs) constructed for specific MCP signaling classes.¹⁵³ Eighty sequences were confidently assigned to the 44H (forty-four helical heptads) class (**Table S1** of Appendix A), which is the main MCP class within the F1 chemotaxis system.⁵⁰ Further sequence similarity analyses revealed a highly conserved group of twelve MCPs exemplified by the Tm14 protein from *T. maritima* (TM0014) and twenty-four MCPs that were sporadically distributed among *Thermotogae* genomes (indication of horizontal gene transfer) remained unassigned. The TM0014-type sequences were all composed of a signaling domain of 36 helical heptads, contained no other domains, and showed a very high degree of similarity. On the other hand, in sequence composition they did not match the previously described 36H MCP signaling class;¹⁵³ therefore, we termed the new class T36H.

One CheW for each class of MCPs. We established that genomes of *Thermotogae* contain two different classes of CheW proteins and MCPs that also belong to two different classes, whereas there was only one CheA protein per genome. Based on this finding we hypothesize that each of the two CheW proteins helps associating MCPs from each of the two classes with the same CheA protein. If so, which CheW is specific for which MCP class? The structure presented herein along with other substantial experimental evidence suggests that CheW1 (TM0701) interacts with the T36H MCP (TM0014).^{61,74} Therefore, if our hypothesis is correct, then 44H MCPs should interact with CheW2. There is no experimental evidence to support this claim, because the CheW2 protein from *T. maritima* has been recalcitrant to recombinant production (data not shown). However, we can offer the following computational evidence in support of this hypothesis. It is a well-established fact in evolutionary molecular biology that if two proteins interact they co-evolve.³² Indeed, CheW1 and T36H proteins appear to be confined to the phylum *Thermotogae*, whereas CheW2 and 44H proteins are widely distributed throughout the prokaryotes.⁵⁰ For example, the sequence similarity score between CheW1 from *T. maritima* and CheW protein from *E. coli* is only 42.4 bits (26% identity), whereas that between CheW2 and the *E. coli* protein is 97.4 bits (37% identity) (**Figure S3** of Appendix A).

If CheW1 interacts with T36H MCPs and CheW2 interacts with 44H MCPs, then class-specific residues in both protein families (CheW1 versus CheW2 and T36H versus 44H) are candidates for the given interaction. Analysis of the multiple sequence alignments constructed from CheW1 and CheW2 sequences (**Figure S4** of Appendix A) and T36H and 44H MCPs (**Figure S5** of Appendix A) revealed several positions in both sets of proteins where correlated mutations have occurred. For example, a position corresponding to Glu12 in TM0701 is 100% conserved as a negative charge in CheW1 orthologs (in one sequence Glu is changed for Asp); however, the same position in CheW2 orthologs is a positive charge (Lys, 100% conserved).

Similarly, the only position where a reciprocal change is seen in the MCP set is Arg131 in TM0014, which is invariably conserved in all T36H MCPs, whereas all 44H MCPs have a negative charge (Glu, 100% conserved). Thus, according to the correlation mutation hypothesis, corresponding positions in CheW and MCP are mutually dependent, e.g. they may interact. The direct interaction between these two residues is not seen in the ternary complex structure because CheW binds at the upper site in the lattice; however, the superposition of CheW onto P5 (**Figure 10C**) predicts that both residues should reside at the periphery of the CheW MCP interface. The CheW-MCP interactions are likely to be dynamic and involve more residues than seen in a single snapshot provided by the crystal structure. The Arg131-Glu12 interaction may participate in recognition of specific receptors by specific adaptors (*vide infra*). Satisfactorily, this finding is further supported by recent NMR studies, where some of the residues showing significant chemical shift changes upon MCP-CheW binding were identified in a very close proximity: rGlu132 (next to rArg131) in Tm14_s and wLeu14 (next to wGlu12) in CheW1.⁷⁴ Other positions with correlated mutations in CheW (wAsp28, wLys36, wAsp38, wGly100, wLys121) and MCP (rThr141, rAsn159, and rGlu162) are also located in the vicinity of residues that showed significant chemical shift changes upon MCP-CheW binding (**Figures S4 and S5** of Appendix A). Most importantly, all predicted CheW-binding sites are located at the MCP tip, in agreement with the 3UR1 structure.

“Top-down” comparative genomic analysis. “Bottom-up” correlated mutation analysis that predicted CheW-MCP interaction sites cannot be used for predicting CheA-MCP interaction sites, because there was only one set of orthologous CheA proteins available for *Thermotogae*. However, we employed a “top-down” approach, where the evolutionary history of the interaction sites prospected from the crystal structure was analyzed for consistency with the correlated mutation hypothesis. For example, if a pair of residues in two proteins is predicted to be an interaction site, they should be either well conserved (e.g. invariable, charge preservation, etc)

or show a correlated mutation pattern. This approach can be equally applied to MCP-CheA and MCP-CheW interactions and we therefore analyzed all four putative binding interfaces revealed in two co-crystal structures. Contact sites in interacting proteins were assigned using a protein cutoff scanning technique¹⁹³ and are shown in **Table S2** of Appendix A. We then used multiple sequence alignments to trace the evolutionary history of each residue in proposed contacts and analyzed it for consistency with the correlated mutation hypothesis. For example, if both residues in a proposed contact pair are invariably conserved, this is consistent with the correlated mutation hypothesis. If one of the residues in the proposed contact is changing in evolution, but another remains conserved, this is inconsistent with the hypothesis. If both residues change and there is a correlation pattern, this is again consistent; however, if both residues change, but there is no correlated pattern, this is inconsistent with the correlated mutation hypothesis. Summarized results are shown in **Figure 10** in reference to the predicted interfaces. Strikingly, the evolutionary history of all 6 MCP-CheA contact pairs and all 4 MCP-CheW pairs prospected on the top spot of the MCP is inconsistent with the correlated mutation hypothesis. In a similarly striking contrast, all 6 MCP-CheA contact pairs and all 4 MCP-CheW pairs identified at the tip of MCP have evolutionary history fully consistent with the correlated mutation hypothesis. Furthermore, in both MCP-CheW and MCP-CheA interactions at the tip, there were true correlated mutations. While the Arg146 in all MCPs remains fully conserved, its interacting residues in CheW show a correlated mutation pattern. Glu12 and Gly100 are mutually dependent (**Figure S4** of Appendix A): Glu12 in CheW1 becomes Lys in CheW2, and Gly100 in CheW1 becomes Glu in CheW2. These residues are mutually dependent, most likely because they both interact with the positively charged invariable positive charge (Lys146) in MCPs. The pattern of co-variance in MCP-CheA interaction is different, but similarly convincing: rIle135 in Tm14 is mutually dependent with kIle566 and kThr565 in CheA (**Figure S6** of Appendix A). rIle135 and kIle566 are in direct contact within the crystal structure (**Figure 10**).

Taken together, the evolutionary history of contact sites assigned from the co-crystal structures strongly suggests that: (i) both CheW and the P5 domain of CheA bind to the tip of MCP and (ii) MCP contact sites for CheW and CheA binding are not the same, although there is a substantial overlap.

Discussion

Comparing the current higher-resolution structure of the *T. maritima* ternary complex to the lower resolution 3UR1 structure previously published reveals important differences that all stem from modeling of the Tm14_s receptor. Firstly, the 3UR1 structure contains four end-to-end hairpin dimers instead of a tetramer of continuous helices. Secondly, the register of the CheW/P5 binding region is shifted roughly one turn of a helix relative to the current structure. In the higher resolution structure, this placement is certain due to clear side-chain density, whereas in the lower resolution structure, the lack of side chain density prevented an unambiguous positioning of the receptor fragment, which was placed based on the apparent positions of the termini.⁶¹ Finally, the polarity of the helical bundles are switched between the structures, which effectively changes the engagement of the lower binding motif from CheW to CheA P5. This all raises the question as to whether the lower resolution data would be better modeled by the new structure derived from the higher-resolution data. Agreement statistics derived from refinement of both models against the lower resolution data does not significantly distinguish the two models (**Figure S2** in Appendix A). Difference Fourier maps between the experimental amplitudes from the two data sets show the largest peaks at the junctions where the hairpins have unzipped (**Figure S2** in Appendix A). This indicates that the two structures may indeed be different in this region, perhaps due to the change in receptor construct. Nonetheless, we must conclude that the lower resolution data does not distinguish the two models.

There is a growing consensus over the structure of membrane associated bacterial chemotaxis receptor arrays, which appear to be universally based on a hexameric assembly of receptors that form trimers-of-receptor dimers, with the greatest ordering at their membrane distal tips where CheA and CheW bind.^{61,76,141,149,202} Recent electron cryotomographic data combined with the modeling of crystallographic structures suggest that CheA and CheW form ring structures of pseudo-hexagonal symmetry that template the receptors, for at least some states of the arrays.^{61,79} Interdigitated assembly states of overexpressed chemoreceptors, where antiparallel dimers associate through their tips have also been described, but are unlikely to be functional.¹⁴⁸ The “unzipped” tetrameric assembly for an MCP subunit described here has not been previously observed, and may well result from the shortened Tm14_s fragment that contains several N-terminal non-native residues at its termini. A similar Tm14_s fragment has been used in NMR studies, where the subunits form helical hairpins and behave as typical MCP dimers.^{74,194} Nonetheless, it is worth noting that full-length Tm14 is a naturally “soluble” receptor, in that it has no transmembrane region.¹⁵⁸ Computational genomics indicates that CheW1 is specific for this class of receptor (T36H). Soluble MCPs have been observed in other settings, where they have important functions.^{102,56,121,203} In some cases these receptors localize to the receptor arrays, whereas in others, they appear to form cytoplasmic clusters.^{56,102,121,203} Given the relatively modest conformational changes in the tip that allow for the switch between dimeric and tetrameric states, the possibility that some classes of soluble receptors form extended unzipped structures should not be ruled out. Notably, virus membrane fusion proteins undergo large-scale conformational changes where hairpin-like helical structures morph into long extended coiled-coil trimers to mediate membrane fusion and viral entry.²⁰⁴⁻²⁰⁷ For these viral fusion proteins, new helical regions form in addition to the extension of the turns, and overall the changes are much greater than what we observe in the structural swap of Tm14_s, where only a few residues change conformation to accommodate the switch. However, the viral proteins

underscore that there is not a prohibitive thermodynamic barrier to the drastic repacking of coiled-coils that accompanies such changes in oligomeric state.

Despite the unusual tetrameric assembly of Tm14_s, we believe that the interfaces found among CheA, CheW and MCPs in this ternary complex structure are representative of those found in the transmembrane chemosensory arrays. Sequence conservation, as well as the co-evolution of interacting sites strongly supports the groove between subdomain 1 and 2 on both P5 and CheW as being the primary binding location for CheA and CheW on MCPs. The evolutionary analysis only supports the lower, tip-proximal binding site as being the recognition motif for both CheW and P5. Although the current high-resolution structure does not have CheW bound at this position, superposition of CheW with P5 generates an interface that was predicted by computational genomics (**Figure 10**), consistent with the hexagonal symmetry of the native receptor arrays, and anticipated by a large body of additional experimental data. NMR chemical shift perturbations implicate residues on both CheW and Tm14_s that are found within this contact (**Figure 12AD**).^{74,194} Pulsed dipolar ESR experiments of spin-labeled ternary complexes of Tm14_s, CheA and CheW combined with targeted disulfide crosslinking also place CheW at the tip of the receptor in close proximity of the lower binding site.²⁰ In addition, allele-specific mutations of CheW that suppress defective Tsr receptors with mutations near or in the lower binding site map to the CheW binding groove (**Figure 12A**).²⁰⁰ Allele-specific suppressor mutations on two genes can imply that the derived proteins interact with one another. Nonetheless, there are other coupling mechanisms possible between suppressors and hence, it is quite remarkable how closely sets of allelic specific Tsr/CheW suppressor mutations localize to the predicted interface between Tm14_s and CheW (**Figure 12A**). Although mutational studies¹⁹⁸⁻²⁰⁰ and chemical modification / protection^{176,208} also support the subdomain1-2 groove

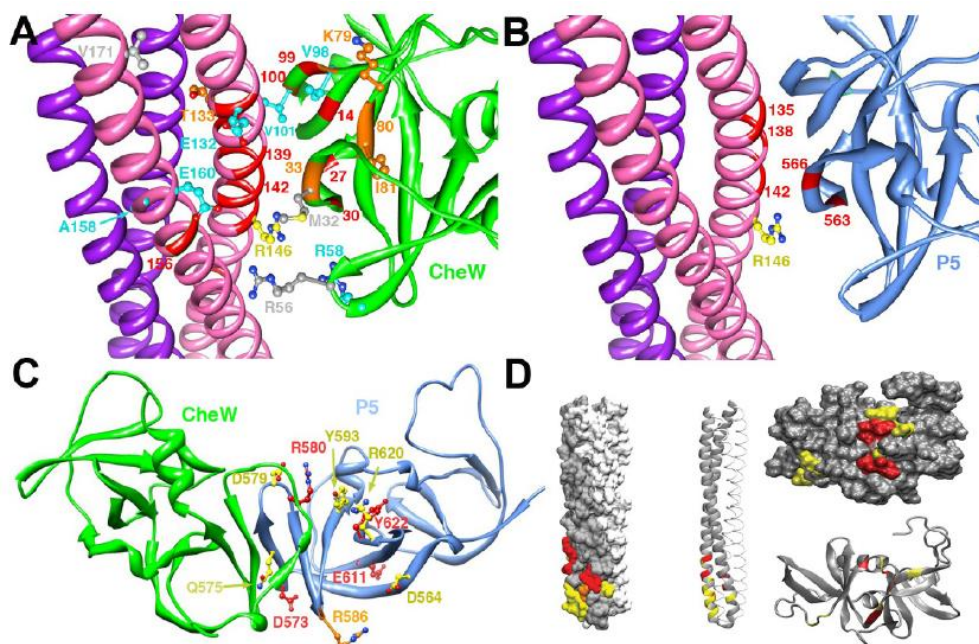


Figure 12. Structure-Function Analysis of Ternary Complex Interfaces. (A) Residue positions known to report on or affect the interaction between CheW and receptors shown on the model of CheW bound to Tm14s based on superpositioning CheW onto P5 bound at the receptor tip (**Figure 10C**). Mutations of CheW residues known to suppress Tsr mutations in *E. coli* are shown as side chains (*T. maritima* numbering). Residue color associates allelic specific suppressors, i.e. mutations at CheW sites that rescue function of mutations at only similar colored sites in the receptor, while at the same time being relatively defective in a wt receptor background (r132 with w98; r133 with w79,w81; r158 with w98, w101; r160 with w57, w58, w101; r171 with w32,w56).²⁰⁰ Orange (and red) ribbons represent CheW mutations or modifications defective in receptor interactions (w27, 30, 31, 32, 33, 35, 80, 81, 98, 101).^{198-200,208} Residues identified as mediating Tm14s-CheW contacts in solution NMR studies shown as red ribbons (w27, 98, 14, 30, 99; R132, 137, 139, 140, 141, 142, 143, 145, 146, 156).⁷⁴ R146 (yellow bonds) resides at the tip of the MCP hairpin and different residue substitutions at this site can produce locked “on” or “off” kinase behavior. (B) Residues found by solution NMR studies to mediate Tm14-P5 interactions are shown as red ribbons (k563, k566; r135, r138, r142).¹⁹⁴ (C) Sites of mutation or modification in P5 subdomain 2 that affect function. Residue positions depicted upon mutation or modification produce chemotaxis defects (red side chains; k573, k586, k580, k611, k622) that curtail CheW binding (orange side chain r586 and rG587 - not shown) or prevent ligand from deactivating kinase (yellow side chains; k575, k579, k564, k593, k620).¹⁷⁶⁻¹⁷⁸ (D) Correlation between bioinformatics and NMR data for predicting CheW interaction with MCP. Left, CheW binding sites on MCP (solvent accessible surface and ribbon representations). Right, MCP binding sites on CheW. Residues predicted by “bottom-up bioinformatics” are in red. Residues identified by NMR are in yellow. Overlapping residues are shown in orange.

on CheW as being the primary MCP binding site (**Figure 12A**), the suppressor studies genetically link this recognition motif directly to the corresponding surface on the MCP tip.²⁰⁰ It should also be noted that there is one set of CheW / Tsr allelic suppressors that map closest to one another when only the upper binding site on Tm14_s is considered (**Figure S7** of Appendix A). Although it is tempting to interpret this data as evidence for functional relevance of the upper binding site, it is more likely that longer-range structural coupling through the receptor propagates the effects of this mutation to CheW binding at the lower position.

Several lines of evidence suggest that the lower binding site on Tm14_s of the current structure represents the primary association mode of CheA with receptors. Specific residues on the receptor tip, where substitutions greatly perturb chemotaxis in *E. coli*, play critical roles in the interface with P5 at the lower site. For example, rAsn139 makes key hydrophilic contacts with partner proteins by hydrogen bonding directly with the main chain of kAle566 and wVal33 (**Figure 10**). Substitutions of the analogous residue in Tsr (Asn381) have dramatic functional effects, with all but a Gly substitution destroying chemotaxis.⁶⁷ In the context of the hexagonal arrays Asn381 is also predicted to mediate receptor trimerization.⁶⁷ Thus, the near essential nature of this residue results from its participation in three distinct interfaces: those with CheA, CheW and two other receptor subunits. In another case, rArg146, which resides at the base of the lower interface (and at the boundary of the tip; **Figure 10**), hydrogen bonds with the main-chain of the connection to subdomain 2 and is close to forming a salt-bridge with kAsp564 (wGlu31). Substitutions to large residues at this site in Tsr (Arg388) produce either “lock-on” (Trp, Tyr) or “lock-off” (Phe, His) kinase activity.¹⁹⁷ Due to its potential to also mediate receptor trimer contacts on the adjacent subunit, these phenotypes also likely result from combined effects at multiple interfaces. Nonetheless, the mutational studies do suggest that alterations in structure at the interfaces resolved by the current structure could be critical for controlling kinase activity.

Despite considerable genetic and biochemical studies of CheA there has been less direct functional data implicating the P5 surface that binds Tm14_s in function.¹⁷⁶⁻¹⁷⁸ This is probably because mutations or modifications at many sites affect P5 structure and CheW binding, and these properties are coupled to receptor interactions.¹⁷⁶⁻¹⁷⁸ Nonetheless, NMR studies that rely on Methyl-TROSY experiments of *Thermotoga* proteins isotopically labeled at select residues identify the same interface defined by the structure and predicted by computational genomics (**Figure 12BD**). Most importantly, the companion paper to this report³⁴ has taken a targeted disulfide cross-linking and mutagenesis (TAM-IDS) approach to define the CheA-receptor contacts in both isolated and cellular chemosensory arrays. These methods were able to distinguish the current interface from the one modeled on the 4.5 Å resolution structure and verify with considerable detail that the interface defined by the higher resolution structure functions in native arrays. Thus, P5 binds at the lower site on receptor tip in a similar orientation to CheW and this at least partially explains the competition of CheA and CheW for overlapping sites on receptors.^{179,209}

Both ternary complex crystal structures contain large rings formed by the P5 and CheW subdomains that have been presumed to template receptor trimers in hexagonal arrays.^{61,79} The ring contact between P5 subdomain 1 and CheW subdomain 2 contact has been characterized by a variety of approaches.^{157,174,176-178,208} The secondary contact that completes the ring structures (subdomain 1 of CheW to subdomain 2 of CheA), has not been previously observed outside of crystal structures; however, mutational data and chemical modification experiments have implicated residues near this interface in function (**Figure 12C**). There are numerous sites in P5 subdomain 2 that when modified produce chemotaxis defects or affect CheW binding; however, many other modifications in subdomain 1 produce similar outcomes.^{177,178} Notably, some Cys-substitutions (and subsequent modification) in subdomain 2 generate phenotypes in which CheA does not deactivate properly with chemoattractant.¹⁷⁷ Such behavior results only

from subdomain 2 substitutions and several reactive sites localize directly to the interface between subdomain 2 and CheW subdomain 1 (**Figure 12C**). Thus, modulation of this ring contact may play an important role in kinase regulation.

Despite our advancing understanding of the overall architecture and interactions within chemosensory arrays, there are many details to be resolved. Array function must involve transitions among different structural states that produce different levels of kinase activity. Recent computational work suggests that hexagonal lattice models may correlate with the active state of CheA, although this remains to be verified.¹⁹⁴ Given what we know about the assembly modes and the large amount of biochemical and genetic data available on the chemotaxis system, can we infer additional interactions not visualized in the current structures? Although caution must be taken in the interpretation of mutational data due to the networked, potentially redundant nature of molecular interactions within the arrays, a few observations deserve note. For example, substitutions or modifications of wArg62 (*E. coli* CheW Arg56) dramatically affect MCP binding and chemotaxis;¹⁹⁸ however, this residue does not directly contact the receptor bundle, despite being oriented towards the C-terminal helix (**Figures 10 and 13**). Furthermore, mutations of Tsr that suppress CheW mutations at Arg62 and residues on the same β 4- β 5 loop map to exposed residues on the C-terminal helix of Tm14_s (**Figure 12A**).²⁰⁰ Notably, Arg62 is also conserved as an Arg on P5 subdomain 2 (kArg586). Thus, it may be possible that in some states of the array, a relative rotation engages the end of subdomain 2 from CheW and/or P5 with the C-terminal helix of the receptor (**Figure 13A**). Indeed, the companion paper demonstrates that targeted disulfide crosslinks between kT565 (*S. typhimurium* E550C) and rI156 (Tsr V398C) increase in the presence of chemoattractant.³⁴ This change in reactivity is consistent with a closer association of subdomain 2 and the receptor C-terminal helix when ligand inhibits CheA.

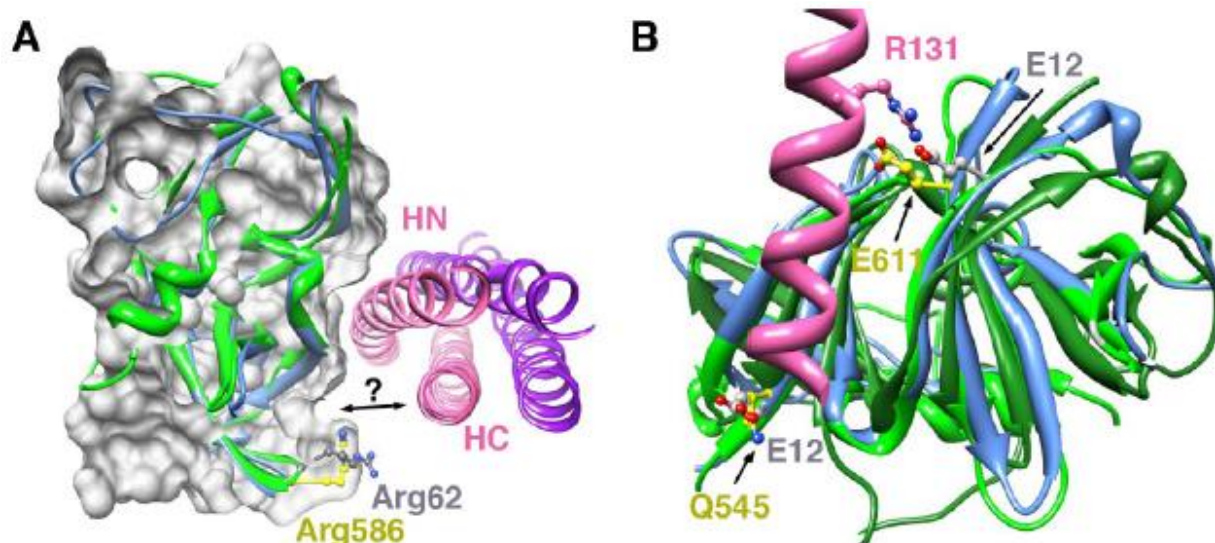


Figure 13. Hypothetical Alternate Interfaces in Chemosensory Clusters. (A) wArg62 (kArg586) is a mutational hotspot on CheW, but does not directly engage the receptor, which binds mainly through the N-terminal helix (HN). A relative rotation of CheW/P5 or receptor could engage subdomain 2 with the C-terminal helix (HC) in some states of the ternary complex. (B) wGlu12 (kQ545) and rArg131 show correlated changes in residue identity. If CheW subdomain is superimposed with P5 subdomain 1 to maintain a similar directionality of helix binding these residues both participate in the interface, but do not contact. However, if subdomain 1 of CheW is superimposed with P5 subdomain 2, wGlu12 corresponds to kGlu611 and would be positioned to salt-bridge with Arg131.

The correlated residue changes at w12 and r131 identified by bioinformatics in the two classes of *T. maritima* CheW/receptor classes also reveal an intriguing structural relationship. The equivalent residue to wGlu12 in P5 is kGln545, but the pseudo symmetric residue on subdomain 2 is kGlu611. kGlu611 makes a direct salt bridge with Arg131, which is the position that undergoes a correlated switch to Glu when w12 changes to Lys. Thus, if the lower binding site were to engage CheW with the N-terminal helix running in the opposite direction, w12 and r131 would salt bridge (**Figure 13B**). Note that this is the orientation that the N-terminal helix takes with respect to CheW at the upper non-conserved site (**Figure 10B**). At least for CheW and Tm14 in solution, PDS and targeted disulfide crosslinking favors the CheW orientation that

aligns like subdomains with P5.²⁰ Nonetheless, in a cellular context it may be possible for a receptor helix to run across the binding groove with both polarities. An inverted arrangement that satisfies the w12-r131 pair would not be compatible with a membrane-associated array, where the receptors all project toward CheA/CheW from the same direction, but such constraints are not present for naturally soluble receptors clusters. Thus, the symmetry of their architectures could be different than those of the membrane arrays and may involve elaborations of the interfaces found in the current structures.

Acknowledgements: We thank the Cornell High Energy Synchrotron Source (CHESS) for access to data collection facilities. We also thank J. S. Parkinson (University of Utah) and F.W. Dahlquist (University of California, Santa Barbara) for helpful discussions regarding the placement of the receptor interaction region relative to CheA/CheW.

Chapter 3: Investigating a Novel Chemoreceptor in *Campylobacter jejuni*

Manuscript Submitted for Review:

A direct-sensing galactose chemoreceptor – a recent innovation of a Cache_1-containing receptor in *Campylobacter jejuni*. **Day CJ, King RM, Shewell LK, Tram G, Hartley-Tassell LE, Fleetwood AD, Zhulin IB, Korolik V.**

Author Contribution:

CJ Day is the first author and was the lead on wet lab experimentation along with RM King, LK Shewell, G Tram, LE Hartley-Tassell, and V Korolik (Korolik Lab, Griffith University). AD Fleetwood and IB Zhulin conceptualized and performed all comparative genomics and bioinformatics analyses. This chapter solely details AD Fleetwood's specific comparative and bioinformatics contributions within the context of generating preliminary results for this collaborative effort. As such, all of the results, figures, and tables from this section were produced by AD Fleetwood.

Abstract

Campylobacter jejuni is a gram-negative member of ϵ -*Proteobacteria* that typically causes self-limiting gastrointestinal illnesses in humans. Foodborne *C. jejuni* infection is often attributed to contaminated poultry products. This pathogen exhibits flagella mediated chemotactic behavior, and two of its chemoreceptors have been characterized. A third chemoreceptor was recently isolated from strains involved in cases where patients were hospitalized due to *C. jejuni* infection. This chemoreceptor was shown to specifically bind galactose via its periplasmic ligand binding domain, and knockout of this chemoreceptor reduced the mutant strain's ability to both mediate chemotaxis toward galactose and also to colonize *in vitro* and *in vivo* models of gastrointestinal infection. Here we show that the *Campylobacter* chemoreceptor for Galactose (CcrG) is a recent innovation of Cache_1 domain-containing chemoreceptors. Through comparative genomics and bioinformatics, we also highlight key differences between the receptor regions in this receptor and closely related receptors in *C. jejuni* in order to identify regions and specific amino acid residues that are potentially involved in the sensing capabilities of three major receptors in *C. jejuni*. While the biological ramifications of a chemoreceptor specifically sensing galactose are currently unclear for *C. jejuni*, the ligand specificity remodeling exhibited by this group of receptors provides a unique opportunity for further study which may have implications for sensory domain adaptation throughout bacterial chemotaxis.

Introduction

We are currently collaborating with the Korolik lab on computationally characterizing three receptors for which they have acquired experimental data. These chemoreceptors were CcmL, which is a multi-ligand sensor which senses several amino acids and other compounds,⁸⁶ CcaA, which is specific for aspartate,²¹⁰ and CcrG, which is specific for galactose.(Korolik *et al.*, submitted for review) CcrG ultimately shows evidence of gene duplication, adaptation of a novel sensory specificity, and entire domain swap events (large scale recombinations). Though many questions remain unanswered, this work shows one example of how new chemoreceptors potentially originate, how they acquire new sensory specificities, and how they might influence behavior or pathogenicity of an organism.

Campylobacter jejuni is the world's leading cause of bacterial gastroenteritis.²¹¹ Through Dr. Korolik's work, one *C. jejuni* chemoreceptor, CcmL, has already been determined to be a requirement for invasion and colonization of chicken and human intestinal epithelium in experimental models of infection.⁸⁶ Dr. Korolik's lab has demonstrated that CcrG contributes to intestinal colonization in experimental models of chicken gastrointestinal infection.(Korolik *et al.*, unpublished data) CcrG is significant not only for this reason, but also because it is a rare receptor that has only been identified in hypervirulent clinical cases requiring hospitalization.(Korolik *et al.*, unpublished data) CcmL or CcmL homologs by contrast are common features in sequenced *Campylobacter* spp. Both understanding the phylogenetic distribution and origin of CcrG are critical for the further characterization of this chemoreceptor. Furthermore, a better understanding of the molecular mechanisms by which chemoreceptor ligand binding domains evolve specificities for and distinguish between different small ligands will greatly enhance our fundamental understanding of signal transduction. Finally, chemoreceptors are druggable, so chemoreceptors potentially involved with pathogenicity like CcrG may one day prove to be valuable therapeutic targets or diagnostic biomarkers.

Results

For the first comparative genomics analysis of CcrG, we established that the *C. jejuni* chemoreceptors characterized by the Korolik Lab share the same chemosensory domain (Cache_1). Cache_1 was not a significant hit when searching CcaA against the Pfam database,²³ and it was a weak or insignificant hit in both CcmL and CcrG (e-9 and e-4 respectively, default significance threshold is e-5) (see **Table 2**). However, through manual multiple sequence alignment, it was clear that these three chemosensory domains were related. Therefore, we employed a more sensitive search method, HHpred,²¹² which detected Cache_1 in the predicted ligand binding regions of all three sequences with >99.7% probability (see **Table 3**).

With the comparative rationale for our analysis solidified, we then investigated the origin and potential mechanism of CcrG's novel ligand sensing domain. Reciprocal BLAST and pairwise alignment methods quickly revealed that the N-terminal half of CcrG is most similar to

Table 2. Pfam Domain Results for *Campylobacter* and *Helicobacter* Chemoreceptors

Gene	Top Hit**	E-Value	Residues
CcrG	Cache_1	1.1e-04	199-263
CcaA	-	-	-
CcmL	Cache_1	1.1e-09	163-231
tlpA	-	-	-
tlpC	-	-	-

Default values* (Raising from E=1 to maximum, E=10, did not improve sensitivity). Pfam 27.0 (March 2013, 14831 families). *This is excluding HAMP and MCPsignal domains, which are common features of methyl-accepting chemotaxis proteins/tlps.

Table 3. HHpred Domain Results for *Campylobacter* and *Helicobacter* Chemoreceptors

Gene	Top Hit**	Probability	Reported Residues	Actual Residues*
CcrG*	Cache_1	99.8	154-237	186-269
CcaA*	Cache_1	99.8	150-233	181-264
CcmL*	Cache_1	99.7	121-192	163-234
tlpA	Cache_1	99.7	118-197	118-197
tlpC	Cache_1	99.7	119-197	119-197

Default settings. Database queried: CDD_19Feb14. *For improved accuracy, ligand binding domains only were assayed in these cases. As such, reported residues must be transformed from the starting residue (32, 31, and 42 respectively). *In both cases, this is excluding HAMP and MCPsignal domains, which are common features of methyl-accepting chemotaxis proteins/tlps.

CcaA (35% amino acid identity, which is substantial within a ligand-binding domain that can be hypervariable), while the c-terminal portion including the signaling domain was **90%** identical to CcmL. This is highly suggestive of a domain swap event, which may have been the result of recombination initiated by the tremendous homology of the chemoreceptor signaling domains (see **Figure 14**).

While these preliminary analyses gave us a local view from the perspective of *C. jejuni*, we needed further phylogenetic depth of comparison to place CcrG in a broader context. To do this, we constructed phylogenetic trees to assess the relationships between representative Cache domain containing chemoreceptors from sequenced ϵ -Proteobacteria: *C. jejuni*, *C. coli*, *C. upsaliensis*, and *Helicobacter spp.*. This analysis was necessary to show that CcrG was not more closely related to any other chemoreceptors than it was to CcaA and CcmL. In order to do so, multiple trees were necessary. These trees used the ligand binding domain, signaling

domain, and full length sequences as separate queries. Unsurprisingly, the LBD tree showed CcrG branching with CcaA, and the signaling domain tree unequivocally showed CcrG branching with CcmL. The full length tree is shown in **Figure 15** in order to place the chemoreceptors that were studied in this analysis in the context of *Campylobacter* and *Helicobacter*. This figure also shows that full length CcrG clusters tightly and confidently within a

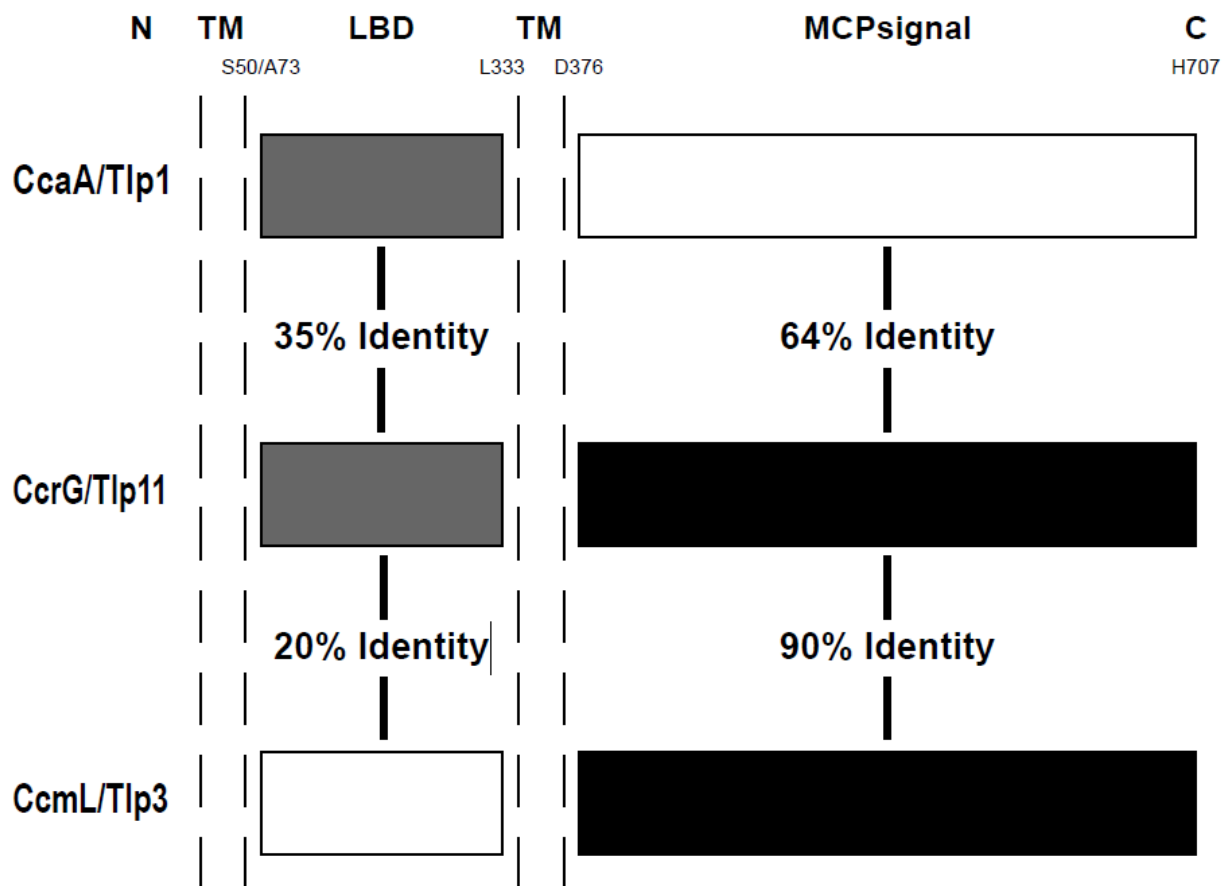


Figure 14: Pairwise Comparisons of CcrG/Tlp11 to Aspartate and Multi-Ligand Receptors in C.

jejuni. BLAST¹⁰⁷, multiple sequence alignment with MAFFT I-ins-i,¹³ and pairwise alignment in Jalview²¹³ were utilized to produce these results.

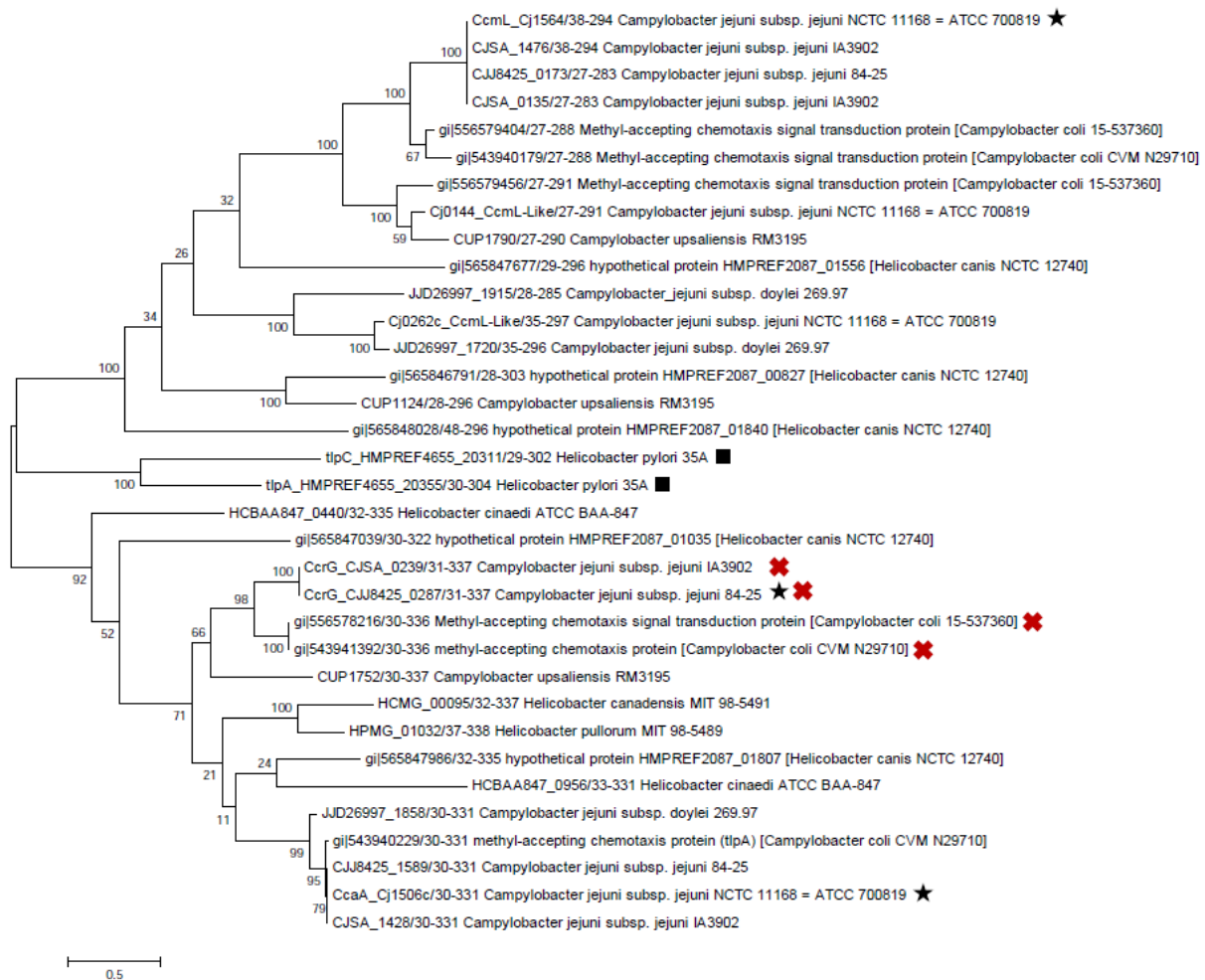


Figure 15: Maximum Likelihood Tree of Campylobacter and Helicobacter Chemoreceptors. This is a maximum likelihood phylogenetic tree of the ligand binding domain regions of the best BLAST¹⁰⁷ hits to the CcrG ligand binding domain. This shows that CcrG and CcaA ligand binding domains are significantly more closely related to one another than they are to CcmL. This also puts the *Campylobacter* chemoreceptors in the context of experimentally investigated *Helicobacter pylori* chemoreceptors, indicating that they have diverged significantly despite sharing the same sensory domain (HHpred predictions support this as well). Finally, a red “X” indicates a hypervirulent or multi-drug resistant strain or clinical isolate, correlating the presence of CcrG with enhanced pathogenicity (though no causality can be inferred from this data alone).

clade enriched for hypervirulent pathogens, bolstering the connection between pathogenicity and the presence of this chemoreceptor and potentially enhancing the impact that the emergence of CcrG has had on the lifestyle of these organisms.

Campylobacter Cache_1 Receptors Differentially Align to CcrG LBD. To produce a sequence and secondary structural alignment of the ligand binding domain and assess how binding specificity may have been altered, we selected representatives of CcaA-type and CcmL-type receptors from the prototypical strain *C. jejuni* subsp. *jejuni* NCTC 11168 = ATCC 700819, CcrG from hypervirulent *C. jejuni* subsp. *jejuni* 84-25, and representative sequences from more distantly related organisms (*C. upsaliensis*, *H. canadensis*, and *H. cinaedi*) in order to show conservation beyond *C. jejuni* for CcaA-type ligands. This alignment was then re-aligned with MAFFT LINS-I alongside the closest available crystal structure to CcmL, which comes from a *V. cholerae* chemoreceptor LBD (PDB:3C8C, Chain A).²¹⁴ Actual crystal secondary structure (2D) alongside consensus 2D predictions from Quick2D (Max Planck Institute) for CcmL and CcaA/CcrG were manually mapped to the alignment (**Figure 16**). Quick2D utilizes multiple secondary prediction algorithms, including PSIPRED, JNET, Prof (Ouali), and Prof (Rost).²¹⁵⁻²¹⁷

CcaA/CcrG and homologs' periplasmic Cache_1 domain region were manually clustered according to the results of pairwise sequence alignment (**Figure 16**). Red rectangles indicate alpha helices and yellow arrows are beta sheets. Residues are colored blue according to identity conservation, with darker blue indicating higher percent conservation. The two types of receptors are clearly related as reflected by the structural and sequence conservation at the start and bottom half of the alignment, yet the two types share a variable region that introduces large gaps in the alignment (**Figure 16**). This region is located in the membrane proximal sensory domain of the LBD and may thus affect ligand specificity. Red boxes highlight positions indicated as in contact with co-crystallized amino acid ligand alanine in 3C8C, with the vast majority only engaging in electrostatic interaction with the amino and carboxyl groups of the

ligand (see **Figure 17** for 3C8C structural visualization). Horizontal black bars demarcate the two types of receptors at these points to highlight that these positions are tremendously conserved with either 100% ID or 100% biochemical property for CcmL-type, whereas in CcaA-type these positions are both relatively non-conserved as well as not containing residues conserved in CcmL-type (see discussion).

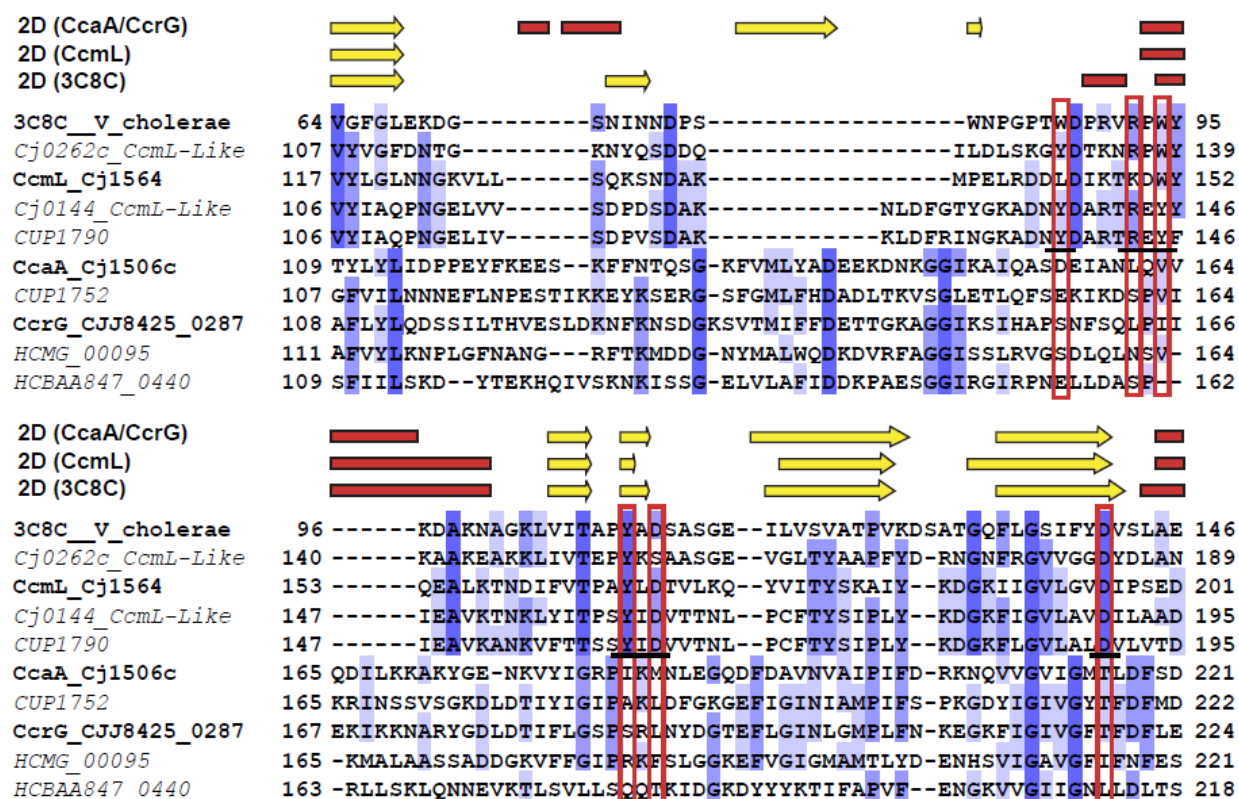


Figure 16. Sequence and 2D Structural Alignment of Cache_1 Chemoreceptors. This clearly shows that there are CcaA type receptors (of which CcrG is one) and there are CcmL type receptors. Though they are all related by the same overall structure (recognized by the Cache_1 domain model in sensitive HHpred searches), they are two distinct lineages that are the result of recent paralogy.

Discussion:

There are four Cache_1 containing methyl-accepting chemotaxis proteins (MCPs) in *Campylobacter jejuni subsp. jejuni* NCTC 11168 = ATCC 700819. Three of these (Cj0144c, Cj0262c, and Cj1564c (CcmL)) have predicted Cache_1 domains through PFAM ²³, whereas the aspartate receptor, CcaA, has a much weaker Cache_1 hit (8.3E-4). The non-aspartate receptors are all of similar length, while the aspartate receptor is 35-41 amino acids longer. Beyond the second transmembrane region, non-aspartate receptors are 100% identical, indicating that they are most likely paralogs that are the result of recent duplications, whereas the aspartate receptor aligns well but has noticeably diverged. To focus on the ligand-binding region, the sequences from all four receptors between the two transmembrane regions were removed and aligned using the MAFFT L-INS-I ¹³ algorithm (see alignment centered on ligand binding residues and the Cache_1 domain in **Figure 16**).

Outside of *Campylobacter*, Cache_1 containing MCPs have been characterized in *Bacillus subtilis* ^{218,219}, *Pseudomonas aeruginosa* ^{220,221}, *Pseudomonas fluorescens* ²²², *Vibrio cholerae* ⁸⁵, and most recently McpU from *Sinorhizobium meliloti*.²²³ These MCPs often appear in paralogous groups (i.e. PctABC in *Pseudomonas aeruginosa* PAO1), and have been experimentally determined to be amino-acid binders. In the case of *Pseudomonas aeruginosa*, out of all three receptors, 18 of 20 amino acids were recognized as chemoattractants, yet none bound aspartate or glutamate. One of these, PctB, was largely specific for glutamine. For *Bacillus subtilis*, another specialized receptor, McpB, was required to recognize asparagine, glutamine, glutamate, and aspartate. For a set of 3 receptors in *Pseudomonas fluorescens*, aspartate again was not sensed though 18 of 20 amino acids were recognized.²²²

One example of Cache_1 receptors from *V. cholerae* was co-crystallized with alanine as the ligand (PDB ID:3C8C ²¹⁴). In this structure, a tyrosine residue and two aspartate residues from the Cache_1 domain are seen to interact with the ligand's amino group. These residues

are highly conserved over a large set of Cache_1 containing proteins (Fleetwood AD and Zhulin IB, unpublished data). In a recent publication from Webb *et al.*, the corresponding aspartate residues in *Sinorhizobium meliloti* McpU were determined to be necessary for proline binding and mutation (even to glutamate) abolished proline recognition. Tryptophan and arginine residues in the “Pre-cache” motif (mentioned by Kawagishi *et al.* in work on Mlp24)⁸⁵ directly N-terminal to the start of Cache_1 form hydrogen-bonds with the carboxyl group, and these residues also show tremendous conservation over a large set of Cache_1 containing sequences (Fleetwood AD and Zhulin IB, unpublished data). This region has been determined to be a globular region that is associated with the Cache domain and should be incorporated in refined domain models (Uphadhyay A and Zhulin IB, manuscript in preparation). Taken as a whole, these five residues recognize the shared structural features of amino acids, leaving the rest of the interacting residues in the binding pocket to potentially serve as determinants of specificity (see **Figure 17**). The non-aspartate Cache_1 receptors show conservation of identity or biochemical properties with all five of these residues, whereas the aspartate residue does not (see **Figure 16**).

The two conserved aspartate residues warrant additional consideration when considering the general inability (or need for a specialized variant) of characterized multi-amino acid binding Cache_1 receptors to bind negatively charged residues and their cognate polar residues. It is possible that electrostatic repulsion excludes aspartate from the binding pocket (whereas the longer sidechain of glutamate might allow for greater charge separation.) When considering other chemotaxis proteins and systems, the aspartate receptor in *E. coli* (Tar) is just one of 5 MCPs, yet this receptor is widely distributed throughout diverse bacterial phyla and is a major driver for motility in these organisms. Aspartate sensing is critical to many bacteria, but in these cases it is sensed through a 4-helical bundle domain, not Cache. We hypothesize that the Cache_1 domain is most suited to bind non-negatively charged amino acids, but through

duplication and neo-functionalization, paralogs (like CcaA) have arisen to compensate. However, in order to sense ligands like aspartate, substantial remodeling of the binding pocket must occur both within specific residues of the Cache_1 domain and also significantly in the structures and loops directly N-terminal to Cache_1.

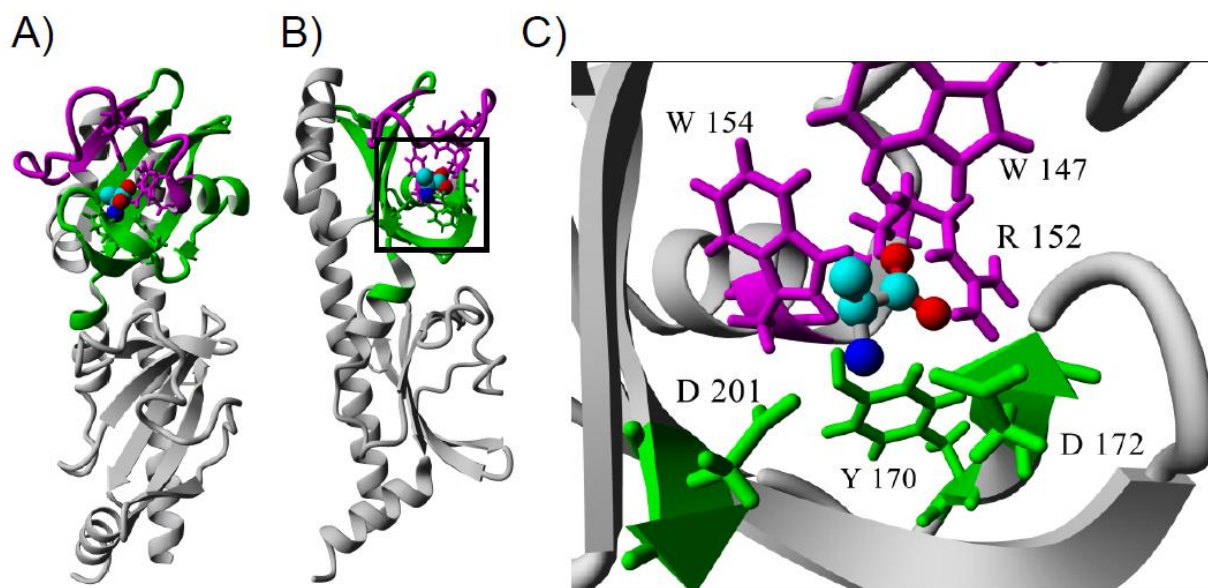


Figure 17: Structural Visualization of LBD with Implications for Amino Acid Specificity Change.

There are two distinct regions of ligand binding residues (a conserved region in green and a variable region in magenta). A co-crystallized amino acid, alanine, is rendered in ball and stick representation (light blue: carbon, blue: nitrogen, red: oxygen). Residues from both regions contribute to ligand recognition in the *Vibrio* crystal structure. Conserved residues (green) visualized above are highly conserved within CcmL-type receptors but show less conservation and even avoidance of the biochemical properties of the conserved residues in CcaA/CcrG. The “remodeling” of these sensory residues may explain the difference between aspartate, galactose, and multi-ligand/multi-amino acid sensing.

Further evidence to the uniqueness of CcrG (and CcaA) can be seen through phylogenetic approaches. To accomplish this, all five Cache LBDs from *C. jejuni* receptors were aligned with a large, phylogenetically diverse set of Cache_1 containing ligand binding regions from other MCPs (data not shown). Then, a maximum likelihood phylogenetic tree was constructed. In agreement with domain architecture and alignment analyses, CcmL-type receptors are more closely related to sequences from *other phyla* (e.g. Spirochetes) than to CcaA and CcrG. A BLAST-based approach also confirmed that CcrG in particular has few sequences that can readily be identified as orthologs, possibly none of which are outside of *Campylobacteriaceae* (data not shown). However, a much more extensive analysis must be conducted to validate that divergent homologs are not orthologs before more can be said.

In sum, many changes in sequence and structure were required for the Cache domain to discriminate negatively charged ligands like aspartate or sugars like galactose. First and foremost, the conserved Cache_1 aspartate residues that are implicated in sensing multiple ligands in CcmL type receptors must be mutated, and in CcaA, these residues have indeed changed to methionine and threonine (M186 and T216). The conserved tyrosine has also changed to isoleucine (I184), so the biochemical properties of this region have changed dramatically (and no longer contain a negative charge). This region of Cache_1 aligns extremely well across all four receptors, and both the observed secondary structure in 3C8C from *Vibrio cholerae* the predicted secondary structure of CcaA are also in agreement. Therefore, these residues are prime targets for mutational studies to probe aspartate sensing in this protein. However, the remodeling of the same region in CcrG is too extensive to pinpoint how aspartate sensing was adapted to galactose sensing without further experimentation. Future work in characterizing the sensory capabilities of *C. jejuni* will expand our understanding of how this organism behaves, as well as having profound implications for sensory domains in signal transduction in chemotaxis and beyond.

Chapter 4: Assigning Chemoreceptors to Pathways in *Pseudomonas aeruginosa*

Manuscript:

A Genomic Approach for Disentangling Multitudinous Chemoreceptors from Multiple Chemotaxis System Pathways in *Pseudomonas aeruginosa* PAO1 and Beyond. **Aaron D Fleetwood and Igor B Zhulin.**

Author Contributions:

AD Fleetwood was responsible for conceptualizing, conducting and overseeing all analyses in this work with Dr. Zhulin. This chapter reflects a full, complete manuscript that is being prepared for submission with *BMC Genomics* as a potential target journal.

*Both Davi Ortega and Jacob Pollack contributed scripting support (16s RNA tree generation, pre-computed sequence retrieval, BLAST) and/or participated in preliminary discussions.

Abstract

Methyl-accepting Chemotaxis Proteins (MCPs or chemoreceptors) are the myriad sensors that prokaryotes employ to relay physicochemical signals which influence and control motility. Chemotaxis has, like many biological phenomena, been best characterized in simple systems like *E. coli*, *S. enterica*, and *B. subtilis*, in which a handful of receptors and only one chemotaxis system are present. However, in several important human pathogens, including *P. aeruginosa* and *V. cholerae*, there are multiple chemotaxis systems (4 and 3) and a multitude of receptors (26 and 45), creating a tangled combinatorial nightmare intractable to *a priori* assignment with current experimental methodologies (e.g. co-localization or chemotaxis phenotype linked mutant studies). As more microbes have been completely sequenced, examples of complex systems like these are in the majority, posing a substantial barrier to deeper characterization of chemotaxis and related signal transduction systems. Therefore, in this work we propose a novel genomic approach to call attention to this issue, aide experimentalists, and expand our understanding of “chemotactically complex” pathogens. In order to *in silico* assign receptors to pathways, this method leverages general features of microbial genome biology using gene neighborhood analyses, nuances of structural biology unique to chemotaxis (helical heptad class assignment and phylogenetic analysis of conserved domains), and phylogenetic profiling. We selected *P. aeruginosa* PAO1 as a test organism, as it is well studied, significantly impacts human health, and has chemotaxis components linked to virulence and pathogenicity. Despite being one of the most highly studied organisms in any kingdom of life, 13 of the 26 chemoreceptors have yet to be experimentally investigated and 16 of the 26 chemoreceptors have yet to be concretely linked to a specific chemotaxis system.

Introduction

Complex Chemotaxis Systems and Receptors. Chemotaxis is a specialized two-component signal transduction system that couples sensory information from chemoreceptors (Methyl-accepting chemotaxis proteins or MCPs) to a histidine kinase, CheA, resulting in a variety of outputs traditionally linked to flagellar motility.²²⁴ Much of the fundamental knowledge of chemotaxis comes from studying flagellar systems in *Escherichia coli*, *Salmonella spp.*, *Bacillus subtilis*, *Campylobacter jejuni*, and *Helicobacter pylori*.^{87,225-227} These organisms only possess one chemotaxis system. However, organisms with multiple chemotaxis systems (indicated by the presence of multiple CheA homologs) outnumber those with only one.⁵⁰ In these cases, the coupling of chemoreceptors to specific chemotaxis systems is anything but straightforward. Core chemotaxis genes composing these systems tend to cluster together in the genome, and occasionally these gene clusters contain a chemoreceptor in the operon as well, making limited wet lab characterization that is system-driven possible. However, the vast majority of chemoreceptors are located outside of these tidy gene clusters, often occurring as genomic orphans. As a result, these receptors are refractory to experimental methods, requiring months of time- and resource-intensive trial and error labor that are not guaranteed to produce results.

State of the Art in Receptor to System Assignment. Despite the difficulties, several organisms with multiple chemotaxis systems have been investigated. One example, *Rhodobacter sphaeroides*, is a model system that illustrates the difficulty of assigning chemoreceptors to CheAs.¹⁰² Only two chemoreceptors out of eleven (McpG and TlpC) have been co-localized and rigorously linked to specific CheA and CheW homologs, in each case requiring numerous mutant strains to rule systems in or out.^{56,103} While other techniques for sorting chemotaxis interactions have been explored in *Rhodobacter*, chemoreceptor system assignment has not been further addressed.^{104,105} In *Pseudomonas aeruginosa* PAO1, aerotaxis chemoreceptors were determined to be dependent on 1 out of 4 potential chemotaxis systems, which required not only deletion-

insertion mutations of all five chemotaxis clusters, but also functional characterization of their sensory specificity (known only for a handful of receptors and accompanied by its own plethora of experimental challenges).¹¹⁷ A final example, *Sinorhizobium meliloti*, has two chemotaxis systems and nine chemoreceptors, yet localization of these receptors required 33 distinct mutant strains and were not pathway specific, as the second system was not included in the analysis.^{124,125}

Bioinformatics and Comparative Genomics Role in Complex Chemotaxis Systems. In *silico* methods have already played a critical role in further elucidating experimental observations for chemoreceptors. However, experimental observations (confocal microscopy and cryo-EM data first showed that chemoreceptors differentially localize to universal, polar membrane-associated clusters.^{76,141,228} Large scale protein sequence analysis of chemoreceptor sequences linked the distance from the inner membrane to the chemotaxis chemoreceptor:CheA:CheW complex by classifying receptor lengths based on detecting helical heptad repeats.¹⁵³ At the systems level, chemoreceptors from *Cyanobacteria* served as one of the first model systems to explore chemoreceptors in their genomic context within multiple chemotaxis systems.¹²⁸ Later, Weis *et al.* used comparative genomics to predict the general function of multiple chemotaxis systems in 3 *Geobacter* spp., in which they observed chemoreceptors from different helical heptad classes and suggested this might diminish unwanted crosstalk between receptors from different systems.¹⁰⁶ In 2010, Wuichet *et al.* conducted a large scale genomic analysis of chemotaxis systems which, aided by recent increases in the number of completely sequenced genomes, was able to identify a large extent to which chemotaxis system variants are conserved and dispersed throughout the known prokaryotic world. By placing these systems within a broader genomic context, clues to their roles and relative importance within organisms can be decoded. Most recently, and concurrent with this work, a large scale experimental study in *Myxococcus xanthus* attempted to reconcile 8 chemotaxis systems and 21 chemoreceptors (seven of eight

operons containing a chemoreceptor), where they showed that receptors and chemotaxis systems that are phylogenetically related (within *M. xanthus*) co-localize experimentally and interact within three major groups (without disentangling receptor:CheA pairs within the larger groups).¹²³

Pseudomonas Chemotaxis. *Pseudomonas aeruginosa* PAO1 is an opportunistic human pathogen that causes a significant number of complicated respiratory and medical device (e.g. catheter) associated infections.¹⁰⁹ Both multi-drug resistance and robust virulence factors contribute to *P. aeruginosa* posing a substantial threat to the immune-compromised and those suffering from cystic fibrosis.¹⁰⁹

Systems. There are four chemotaxis systems and 26 chemoreceptors in *Pseudomonas aeruginosa* PAO1.²²⁹ Che cluster I (PA1456-PA1464), an F6 flagellar system, was described to be essential for chemotaxis as CheY and CheZ mutants were deficient in flagellar-mediated motility/chemotaxis.²³⁰ Che cluster 5 (PA3348-PA3349), containing a CheV and CheR protein, was linked to Che cluster I soon after.¹¹² Che cluster II (PA0173-PA0180), a divergent F7 system, was partially characterized and determined to have higher homology with the F7 system in *E. coli* than with other *Pseudomonad* systems, with potentially analogous systems in *Vibrio cholerae* and *Shewanella oneidensis* MR1.¹¹⁰ Che cluster III (PA3702-PA3708) was characterized and renamed WSP and determined to regulate biofilm formation through control of c-di-GMP turnover, and it has since been categorized as an ACF (Alternative Cellular Functioning) system and is associated with the MCP WspA (PA3708).⁹² Finally, Che cluster IV (PA0408-PA0417), the CHP system, is a system in control of twitching motility (a TFP or Type IV Pilus system) and is associated with the PilJ MCP (PA0411).⁹¹

Chemoreceptors and Localization Progress. Three paralogous transmembrane receptors, PctABC, have been determined to mediate taxis towards amino acids.¹¹²⁻¹¹⁵ PctABC localize to the major chemoreceptor cluster at the cell pole. Two cytoplasmic chemoreceptors (BdIA/PA1423

and McpS/PA1930) have been investigated, with the former linked to nitric oxide sensing and biofilm dispersion phenotypes and the latter having been localized to the polar chemotaxis array.^{120,121,231} While the exact relationship is unclear, both soluble receptors appear to be homologs of AerC, a cytoplasmic chemoreceptor shown to link metabolic changes to chemotaxis in *Azospirillum brasilense*.⁵⁷ In order to probe the determinants for localization of receptors from different systems, the highly conserved MCP signaling domain was swapped between the PctABC and WspA receptors, showing that this region was necessary and sufficient to localize each to their wild-type position (WspA, unlike PctABC, is located laterally).¹¹⁶ McpA (PA0180) and McpB (PA0176) were characterized and demonstrated generalized flagellar-motility defects, yet while McpB was determined to preferentially localize with the F7 flagellar system, McpA (which is in the F7 gene cluster alongwith McpB) instead co-localized with the F6 system.¹¹¹ The two systems do not appear to co-localize, and the entire F7 system (along with four chemoreceptors – McpS/PA1930, PA2573, PA2920, and PA4915) has been shown to be induced by RpoS,²³² leading Harwood *et al.* to posit that the polar cluster is remodeled during periods of stress. Both CheB2 (PA0173) from the F7 system and PA2573 have been implicated in pathogenicity in experimental models of infection.^{83,84}

Localization Progress. While great strides have occurred in this field since 2005, only incremental progress has been made in fully characterizing both the F7 system and the uncharacterized MCPs. *P. aeruginosa* chemotaxis was reviewed in 2008, and an even more recent review was published at the time of writing this article, both of which provide more comprehensive overview of *Pseudomonas* chemotaxis.^{233,234} While half of the 26 chemoreceptors in *P. aeruginosa* PAO1 and a few examples from other *Pseudomonads* have had a sensory specificity investigated or have been otherwise studied, only 10 receptors have been localized or experimentally assigned to a pathway (see **Figure 18**).^{118,235,236}

Results

Gene Neighborhood Organization. In bacteria, gene neighborhood analyses can provide a good starting point for characterizing protein systems, complexes and enzymatic pathways due to the tendency for genes to be grouped in operons that are co-transcribed and co-expressed.²³⁷ In a previous study from our lab, the gene neighborhoods of chemotaxis systems from 450 prokaryotic genomes were extensively studied and catalogued.⁵⁰ From this dataset, out of 487 chemotaxis systems (identified by the presence of CheA), there are 302 instances of an MCP either directly in the operon or in a chemotaxis system operon within 3 genes of CheA or a CheA containing operon. Thus, 67% of all chemotaxis operons or systems are paired with an MCP, indicating significant enrichment vis a vis the rest of the genome.

For PAO1, gene neighborhood analysis has been previously exploited to characterize five chemotaxis clusters. In **Figure 18**, we have visualized the core chemotaxis proteins CheA, CheW (CheV), CheB, and CheR from these clusters. 26 chemoreceptors, along with their predicted membrane or cytoplasmic localization are also present. From gene neighborhood analysis alone, only 4 receptors out of 26 have links to a particular system. However, four additional receptors (PctABC and Aer) have also been linked to Che Cluster I (F6). The remainder are genomic orphans with no evidence for assignment, leaving 18 MCPs unassigned.

HMM Analysis for Heptad Class Heptad Length. Wuichet *et al.* noticed a tendency of chemoreceptors from specific heptad classes to associate with chemotaxis systems based upon the chemotaxis system classification scheme designated in that work (see **Table 4**). The sequence of the highly conserved MCP signaling domain (the CheW:CheA:MCP interface), the region directly N-terminal up to the methylation and adaptation region (where CheB/CheR interface), and the C-terminal region can be easily recognized due to a consistent helical heptad repeat. This structural signature is mirrored at the hairpin tip of the chemoreceptor, creating a situation in which an HMM can classify and cluster chemoreceptors by heptad (H) length.¹⁵³

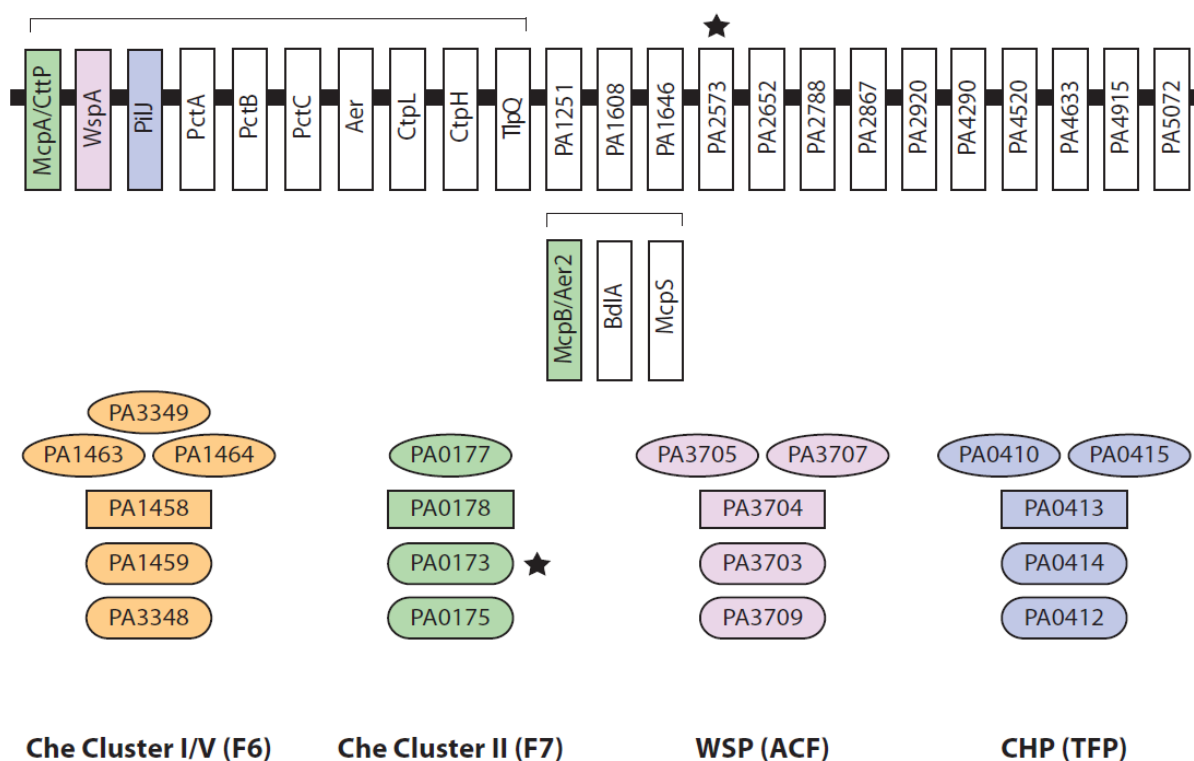


Figure 18. A Priori Chemotaxis Systems and Chemoreceptors in *Pseudomonas aeruginosa*.

Pathways encoded by *P. aeruginosa* PAO1 gene clusters I and V (beige) and cluster II (green) control flagellar motility. Pathways encoded by cluster III (magenta, WSP system) and cluster IV (blue, CHP system) control cyclic di-GMP turnover and Type IV pili motility. Chemoreceptors are shown as long vertical rectangles. Characterized/named chemoreceptors grouped by horizontal brackets. Chemoreceptors linked to particular pathways through gene neighborhood analysis are shown in corresponding color. Key components for each chemotaxis pathway are shown: CheW/CheV adaptor proteins (ovals), CheA histidine kinases (horizontal rectangles), CheB methylesterases and CheR methyltransferases (rounded rectangles). Proteins implicated in pathogenesis are marked by an asterisk.^{83,84}

Chemoreceptors with different heptad classes are present in a number of organisms including *Geobacter spp.*, where it was suggested that the differing lengths might diminish unwanted crosstalk between chemotaxis systems.¹⁰⁶ This view of the relationship of

chemoreceptor length to chemotaxis complex array is biologically relevant, as cryo-EM images of chemotaxis arrays have shown the CheA layer of the array as a line that is roughly equidistant from the periplasmic membrane.⁷⁶ This indicates that the receptors which span from membrane to CheA within the same array are the same length. That said, we make two assumptions for this analysis: 1) the chemoreceptor belongs to a chemotaxis array and 2) that the chemoreceptor has canonical transmembrane receptor architecture of a ligand binding domain bounded by two transmembrane domains. This concept guides our second analysis, in which we investigate the correlation of heptad lengths of PAO1 chemoreceptors to the four systems and their known associations.

To examine whether or not the heptad class association with specific systems from **Table 4** continue to hold with the massive increase in sequences generated between 2010 and now, we searched the Mist 2.2 database (2,756 complete genomes) for all organisms with only

Table 4. Chemotaxis Systems, System Designation, and Heptad Class Association.

Chemotaxis System	Chemotaxis System	Associated Heptad
Cluster	Designation	Class
CheI/CheV(5)	F6	?
CheII	F7-Divergent	36H
CheIII	Tfp	40H
CheIV	Acf	40H

This table summarizes trends of MCP enrichment within specific chemotaxis operons that were noted by Wuichet *et al.*⁵⁰ No trend was reported for the F6 system, likely because major representatives of the F6 system like *P. aeruginosa*'s F6 system do not contain any chemoreceptors. This necessitated our analysis in **Figure 19**.

one chemotaxis system and asked one question: what was the heptad class of the receptors in these genomes (**Figure 19** shows results for *P. aeruginosa* PAO1)? The results of the heptad class distribution of chemoreceptors from organisms with only a single chemotaxis system (F6, F7, ACF, or TFP) are presented in **Figure 20**.

By this analysis, we predict the 36H receptor to preferentially associate with the F7 system. This is consistent with gene neighborhood analysis results, because the 36H receptor is in the F7 operon (Che cluster II). However, there are 21 40H chemoreceptors which are strongly associated with the F6 system (Che cluster I). Two of these 40H receptors (PilJ and WspA), were previously assigned (see **Figure 18**) due to their presence in the TFP and ACF operons respectively. While we cannot rule out crosstalk between systems that prefer 40H receptors based solely on the

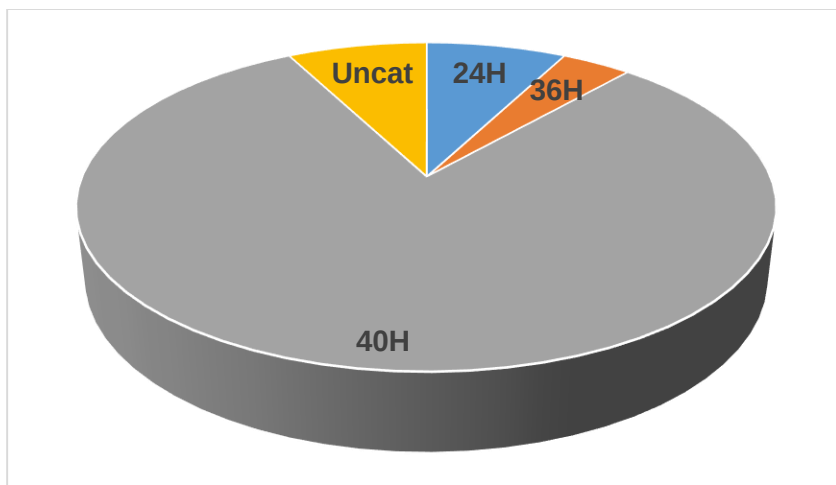


Figure 19. Heptad Class Distribution for *Pseudomonas aeruginosa* PAO1 (26 MCPs). In order to obtain the heptad length information for the 26 PAO1 receptors, the sequences were searched with Profile-HMMs for all known heptad classes. Uncategorized (Uncat) MCPs do not conform to the canonical TM-LBD-TM-HAMP-MCPsignal domain architecture with symmetric heptad profiles reflecting at the MCP tip as previously described. Distribution: 40H: 21 receptors, 36H: 1 receptor, 24H: 2 receptors, Uncategorized: 2 receptors.

criteria of heptad class, to the best of our knowledge there is no evidence from the literature that the TFP or ACF systems interact with MCPs other than those within their operon. Furthermore, Harwood *et al.* swapped the C-termini of 40H receptors (PctABC with WspA) and found that they localize separately, providing experimental evidence against their inclusion into the polar F6 chemotaxis array.¹¹⁶ Thus, 19 40H chemoreceptors are linked to the F6 system, leaving only 3 receptors (2 24H and one uncategorized) with no pathway links.

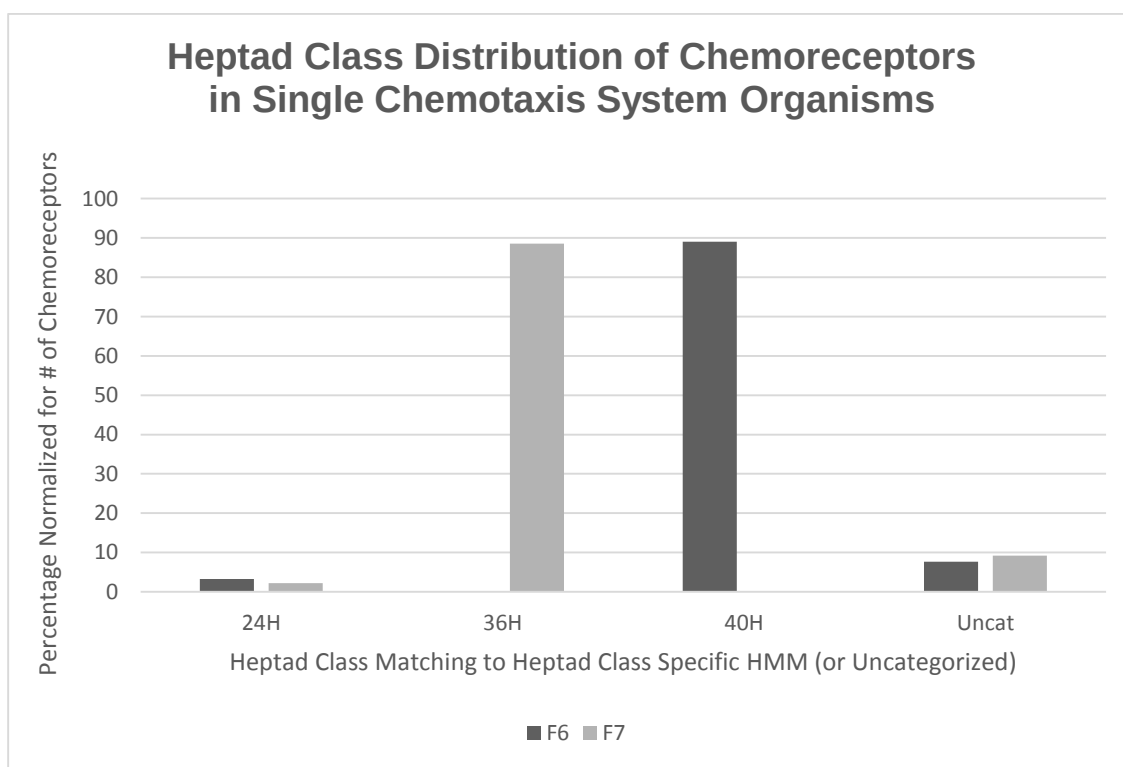


Figure 20. Helical Heptad Length of Chemoreceptors in 1 System Organisms. Strikingly, the trends continue to hold, especially in the case of F7 exclusive systems, which preferentially contain 36H class receptors vs. 40H class receptors (one exception out of >2000 sequences). Satisfactorily, F6 exclusive systems do not contain a single 36H receptor. TFP and ACF, which are highly diverse systems that do not modulate flagellar-mediated motility, do not show heptad exclusivity. Interestingly, 24H receptors are present in both F6 and F7 exclusive systems, indicating that they may potentially interact with both systems.

Genomic Subtraction Analysis. After looking at the chemotaxis system gene neighborhoods, as well as the helical heptad class of the chemoreceptors, we decided to employ phylogenetic profiling to assist our search. For this analysis, we chose the order *Pseudomonadales* for our dataset, because it provides a cohort of the organisms closest to PAO1 while also allowing enough evolutionary distance within which major changes to chemotaxis system organization could have occurred. This group also contains numerous completely sequenced organisms. In **Figure 21**, a phylogenetic tree of 16S RNA from those genomes completely sequenced from order *Pseudomonadales* as of May 2013 is paired with the chemotaxis systems present in each genome (as detected by profile HMM search).

Our logic for the analysis of this data was as follows: if receptors in PAO1 are specific to a chemotaxis system, then orthologous receptors should be present in closely related organisms with CheA orthologs for that system and absent when that CheA is not detected. Interestingly, in all genomes from our dataset except *Moraxella cattarhalis* and *Pseudomonas stutzeri* ATCC 17588 = LMG 11199, the TFP system is present (**Figure 21**). This system appears to have been present in the common ancestor of *Pseudomonadales*, and its retention belies its importance to all of the organisms in this group aside from *Moraxella*. The chemoreceptor in the PAO1 operon with this system, PilJ, has been experimentally characterized,⁹¹ and we observed orthologs in all of these organisms (including the *P. stutzeri* lacking the system). On manual inspection of the gene neighborhood of the PilJ ortholog in this organism, the CHP system CheA ortholog (ChpA) has a stop codon in the nucleotide reading frame for the CheA that appears to have created a non-functional gene/pseudogene (which was undetected as a CheA homolog by any CheA-specific HMM). This was verified against the nucleotide sequence of the other *P. stutzeri* (str. RCH2) from our dataset, which does not contain a stop codon and whose ORF contains a full length CheA protein. *Acinetobacter* spp., including *A. baumannii*, an emerging nosocomial pathogen,²³⁸ only possess one MCP (PilJ) and one TFP chemotaxis system, illustrating the 1:1

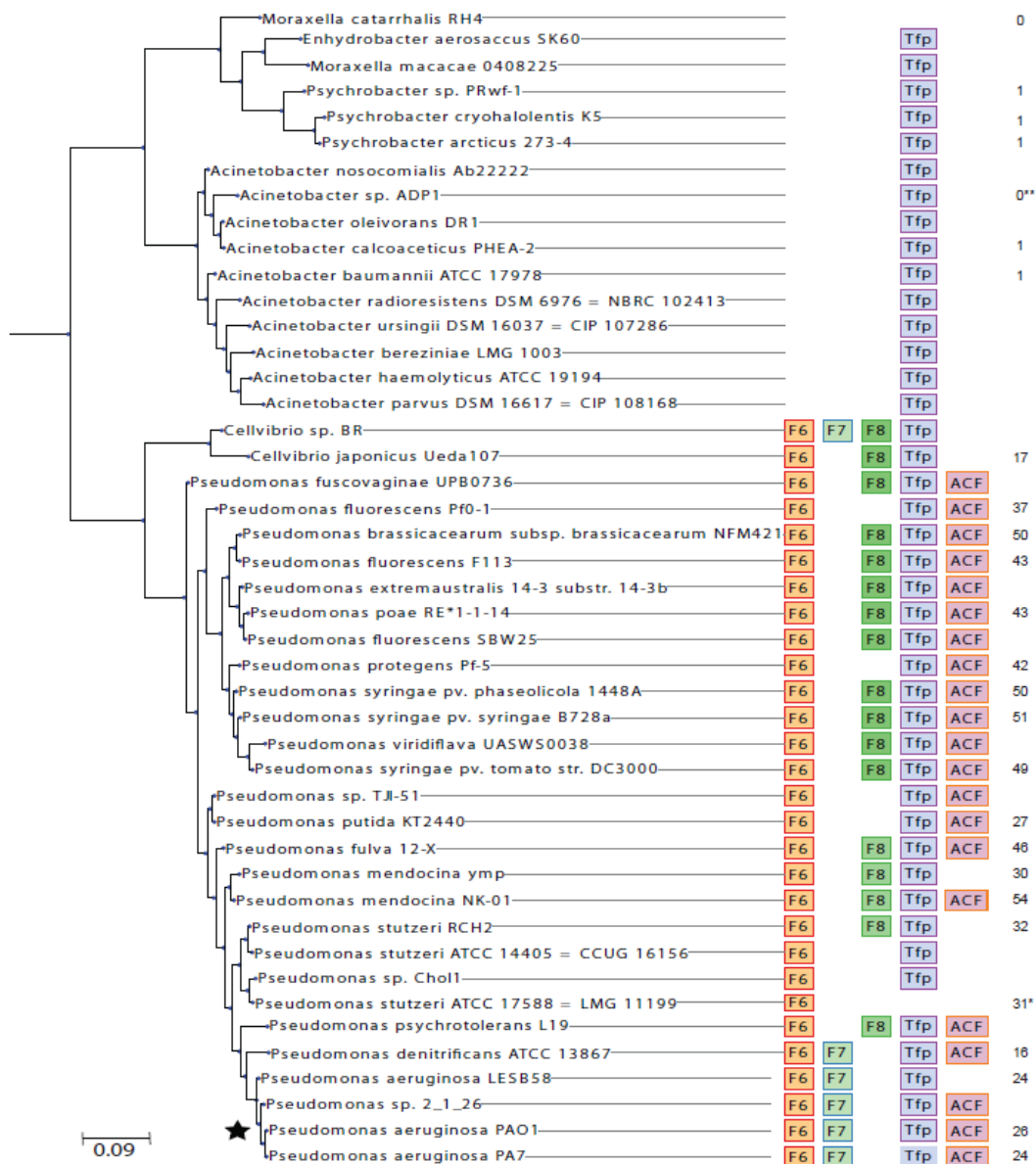


Figure 21. 16S RNA Tree for Order *Pseudomonadales* with Chemotaxis Phylogenetic Profiles. *P. aeruginosa* PAO1 denoted by star. TFP (purple) and ACF (pink) are non-flagellar systems. F6 (orange), F7 (cyan), and F8 (green) are flagellar systems. F8 is not present in PAO1 but is evolutionary related to the F7 system.⁵⁸ Organisms were chosen as representatives at the species level and/or when completely sequenced genomes were available. Number of chemoreceptors detected in complete genomes present in MiST 2.2 database shown on far right. **P. stutzeri* strain with only F6 system yet **more** chemoreceptors than PAO1. **Possible sequencing error in PilJ homolog of an organism where twitching motility has been studied.²³⁹

relationship of this specialized chemotaxis system as a sharp contrast to other cases that are not as simple.

Since the WSP/ACF system has also been experimentally paired with its operon-MCP (WspA), we looked for WspA orthologs in all organisms as well. We found orthologs only in organisms with a predicted ACF system with one exception, a hypervirulent cystic fibrosis epidemic strain, *P. aeruginosa* LESB58.²⁴⁰ Upon manual inspection of the gene neighborhood of the WspA ortholog, the WSP system CheA ortholog (WspE) was detected with a frameshift mutation that rendered the N-terminal half (including the HPT phosphor-transfer domain) completely abolished, while the C-terminal half was intact. Whether or not this mutation in the ACF system (which could affect biofilm regulation) contributes to the hypervirulent phenotype of this strain may be worth further consideration. This may also provide fundamental insight into the contribution of WspE to the ACF system (specifically the HPT domain) if the truncated mutant is expressed lacking a major regulatory domain.

We next observed that all organisms in this dataset from *Pseudomonadales* with flagellar motility systems have the F6 system (**Figure 21**). While all other flagellar organisms possess more than one chemotaxis system, serendipitously, *P. stutzeri* 17588 contained only the F6 system. This created an opportunity for comparison to PAO1, as F6 exclusive receptors could be identified in PAO1. *P. stutzeri*, despite having two fewer chemotaxis systems, contains 31 putative chemoreceptor genes, which is more than PAO1, further supporting its role as the major flagellar system. From this analysis, 17 chemoreceptors had orthologs in *P. stutzeri* (as determined by manual alignment, pairwise alignment, and reciprocal BLAST, data not shown), 15 of which were 40H, one of which was 24H (PA1423), and one of which was uncategorized (PA4290).

Robust Chemoreceptor Ortholog to Chemotaxis System Co-Occurrence. The next step that we took to link remaining receptors to pathways was to identify high-confidence orthologs of

these receptors at deeper taxonomic levels in order to expand our phylogenetic profile of which systems and chemoreceptors co-occur. The initial input into our robust, manually curated workflow was a BLASTp query. Multiple sequence alignments, maximum likelihood phylogenetic trees, domain architecture prediction, and reciprocal BLAST were all employed to produce a conservative set of orthologous sequences (methods as previously described). A tightly controlled, conservative set was necessary due to the high levels of gene loss, duplication, and horizontal gene transfer demonstrated in chemoreceptors.¹⁵² While, only one receptor lacked a line of evidence for pathway assignment at this step (McpS/PA1930), we included BdlA/PA1423, McpA/CttP/PA0180, and PA4290 as they either lacked canonical transmembrane architecture (precluding them from heptad analysis) or had conflicting experimental and genomic data (McpA). The resulting distributions were then analyzed (data not shown).

PA4290 is a unique chemoreceptor in this set, as it is the only instance of a receptor with a putative ligand binding domain on the c-terminal side of the highly conserved MCP signaling domain. While a search against the PFAM database detected no significant domain hits to this region, a more sensitive domain detection tool, HHpred,²¹² predicted the Cache_3 domain, a known signal transduction and chemoreceptor sensory domain, with greater than 97% probability. In the robust ortholog:system co-occurrence analysis, 47 organisms contain high confidence orthologs, four of which have only one system (all F6). Additionally, only 1 of 47 organisms is missing the F6 system, and that organism contains an F3 system which is phylogenetically related to F6.⁵⁰ Coupled with its earlier presence in the F6 exclusive *P. stutzeri* strain, multiple lines of evidence point to PA4290 interacting with the F6 system over any other.

McpA/CttP (PA0180) orthologs do not occur in any organisms with only one chemotaxis system, and in every case where there are only 2 systems present, the systems are F6 and F7. Out of 54 ortholog-containing organisms, there are only 4 cases where F6 and F7 are not jointly present. It is interesting to note that a divergent homolog of McpA is also present in the same

gene neighborhood of the F8 system in other *Pseudomonadales*. These could not be characterized as high-confidence orthologs as they only share ~27% amino acid identity over the full length (390aa) of the sequence with McpA, whereas most of the other sets of high-confidence predicted orthologs in this work share roughly 60-70% amino acid identity over the same length. While domain architecture, multiple sequence alignment, and gene neighborhood similarities indicate that this receptor could potentially perform a similar function, the mechanism and nature of McpA and its divergent homologs require additional experimental characterization.

Finally, the 24H cytoplasmic AerC homologs, BdlA (PA1423) and McpS (PA1930) are ostensibly the two most difficult to analyze using sequence information. Both have identical domain architecture, yet they only share ~39% amino acid identity over 417aa. While it is tempting to speculate that they are recent paralogs, both the low identity percentage and BLASTp of each against the NCBI NR database suggest different evolutionary lineages instead. PA1423's closest hits are in the *Pseudomonadales*, yet PA1930 hits *Shewanella* and *Vibrio* spp. before other *Pseudomonads* outside of the *aeruginosa* group. While BdlA orthologs appear to be enriched in F7 exclusive organisms, it is also present in multiple system organisms without F7 as well as in the F6 *P. stutzeri* analyzed earlier, suggesting that it can interact with both systems. PA1930 is present in F6, F7, and F3 exclusive organisms, so while no conclusions can be drawn as to exclusivity, it also appears possible that it could interact with both systems.

Phylogenetic Clustering of Pseudomonadales Chemoreceptors by Signaling Domain (Figure 22). As a final step, we manually constructed an alignment of the MCP signaling domain for all 696 predicted chemoreceptors within complete genomes from order *Pseudomonadales* in the MiST database²¹ from the same organisms in our genomic subtraction analysis. The highly conserved MCP signaling domain serves as the interface for interaction with CheA, making it the most likely region to provide MCP:CheA information. Approximately 4 heptad lengths mirrored at the hairpin tip residues constituted a gapless alignment for all 696 sequences. This final alignment region was 59 amino acids in length. A phylogenetic-based

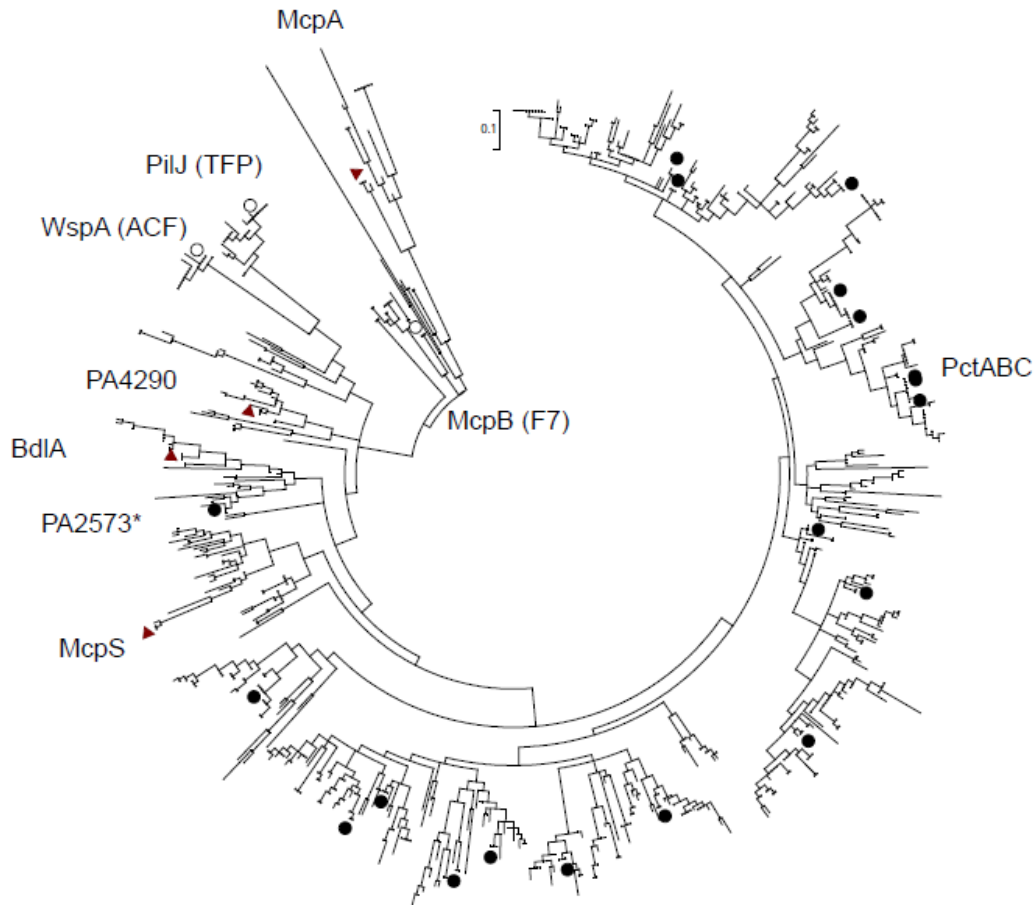


Figure 22. 696 Chemoreceptor Signaling Regions from *Pseudomonadales*. Black circles indicate 40H receptors with multiple lines of evidence for interaction with F6. Open circles indicate receptors both predicted and known to not interact with F6 (WspA, PilJ, McpB). Red diamonds indicate uncategorized or soluble 24H receptors. Asterisk denotes PA2573, which is one orphan chemoreceptor implicated in pathogenicity.⁸⁴

approach was recently used on a limited scale for *Myxococcus xanthus* chemoreceptors (though the MCP set was confined to only *M. xanthus*).¹²³ We then built a maximum likelihood phylogenetic tree from this alignment (see **Figure 22**). We also constructed minimum evolution and neighbor-joining trees using the same alignment to ensure that similar topology was achieved through each method (data not shown). In agreement with our previous findings, TFP

associated chemoreceptor PilJ and ACF associated chemoreceptor WspA both form distinctly separate clades. McpB, which is present in the F7 operon and has been localized with the F7 system also forms a separate clade. McpA, which is uncategorized and has conflicting experimental and genomic data, clusters in the most divergent branch of the tree. All other chemoreceptors form branches that are clearly distinct from the aforementioned groups but related to one another. It appears that this large set contains primarily 40H receptors and 24H receptors, which by this analysis cluster more closely with F6 system receptors. This further supports the notion that F6 is the major system in *Pseudomonadales* and by extension, PAO1.

Discussion:

Since this is a well-studied system, *P. aeruginosa* PAO1 is less of a discovery model and more of a proof of concept than many other microbes might have been. However, we provide the first lines of evidence for numerous orphan receptors associating with any given pathway. We also provide additional lines of genomic data that support experimental work and the conclusions reached in previous work. We were able to assign all of the chemoreceptors to pathways, though cytoplasmic receptors and McpA still have lines of evidence that support interaction with F6 and F7 (leaning in favor of F6). Residue level analysis may provide better resolution for these cases. TFP is the ancestral system for *Pseudomonadales*, and the F6 system is the predominant flagellar system that lays claim to the lion's share of chemoreceptors. While the assumption that the first chemotaxis system characterized was the major one, there is now genomic evidence for this, which in and of itself will greatly expedite the characterization of the remaining PAO1 receptors.

Additionally, we observed several trends that may provide insight into *Pseudomonad* biology. First, there are systems that are specific to the *P. aeruginosa* clade (e.g. F7) and to the *P. syringae* clade (F8). *P. aeruginosa* is an opportunistic human, animal, and plant pathogen, while *P. syringae* is a well-known specialist plant pathogen. The receptors coded within these

corresponding operons may be key components for niche colonization. Even within the aeruginosa group, two strains (PA7 and LES) show markedly different behaviors (PA7 is a non-infectious environmental sample and LES stands for Liverpool Epidemic Strain).^{240,241} PA7 and *Pseudomonas denitrificans* (not known to cause disease to the best of our knowledge) are missing several chemoreceptors that the other two more infectious strains possess, including PA2573 and PA4915 which were either implicated in pathogenicity and/or upregulated by RpoS stress. Chemoreceptors serve as the “GPS” for the cell, biasing its movement towards favorable environments and away from those that are unfavorable. However, the chemoreceptor suites can be highly variable, as they are subject to constant duplication, gene loss, and horizontal gene transfer. Therefore, these receptors should not be overlooked as potential virulence factors, as their role in shaping the movement and flow of organisms between environments (even within the body) may lead to unintended pathogenicity consequences. We observed a similar correlation between receptor loss and pathogenicity in the B2 pathotype of *E. coli*, where two key chemoreceptors were lost in an ancestral event and the resulting clade was highly enriched in extra-intestinal pathogens (**Chapter 5**).²⁴²

The bulk of receptors are predicted to preferentially associate with the major F6 system (**Figure 23**), and the more specialized systems have very limited chemoreceptor suites, often limited to one receptor contained within the operon. It will be interesting to see if this trend holds in other organisms with complicated chemotaxis pathways. While CHP and WSP have been studied extensively, the small specialized flagellar systems have not, and these may hold the key to understanding the dynamics of these complex chemotaxis systems and their role in the behavior of these organisms. Moine *et al.* recently showed in *M. xanthus* that related chemotaxis systems and chemoreceptors co-localize and may form complex, interconnected groups, so a similar scenario may occur in *Pseudomonadales* between the F6, F7, and F8 systems.¹²³ Both the mechanism of their co-existence and their ability to differentiate closely related homologous

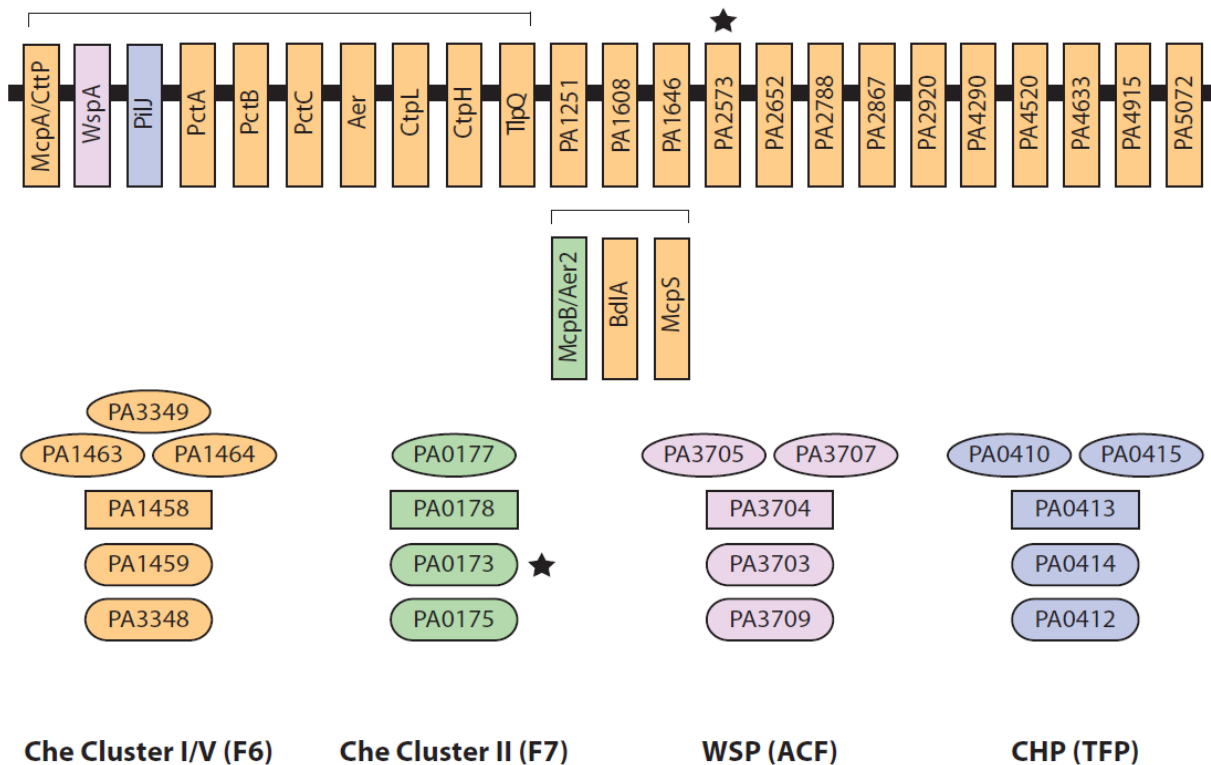


Figure 23. New Look: Multiple Chemotaxis Systems in PAO1 Disentangled. All receptors now have genomic, phylogenetic, or other comparative data linking them to at least one specific system. Our view of the chemotaxis landscape agrees with the experimental consensus that the F6 Che Cluster I system is the major flagellar system. However, it has not yet been shown that the vast majority of uncharacterized, mono-cistronic chemoreceptors are predicted to interact with this system, and the presence of another flagellar system (F7 Che Cluster II) had previously complicated this picture. The fact that both specialized flagellar systems and orphan receptors contribute to pathogenicity (in addition to previously known systems like WSP and CHP) has expanded the impact of chemotaxis as a source of virulence and avenue for therapeutic targeting. Our analysis and methods provide the first steps towards making whole-scale systems characterization of multiple chemotaxis systems and multiple chemoreceptors an attainable reality.

components will be the focus of future work in this area. If these specialized flagellar systems are enriched in pathogens vs. non-pathogens, they, like the CHP and WSP systems, along with their

associated chemoreceptors may provide excellent candidates for specific, novel classes of antibiotics with lower side effect profiles against commensal flora.

In silico assessment of chemotaxis components before experimental investigation will save labs planning to do future work in chemotaxis a tremendous amount of time and resources, and may serve as a starting point for less studied and as yet wholly unstudied chemotaxis systems. These complicated cases comprise over half of sequenced chemotactic organisms, and there are over 58,000 non-redundant chemoreceptors in the current RefSeq database.⁴⁹ Thus, this method stands to make a significant impact on the global characterization of chemoreceptors across a wide variety of plant, animal, and human commensals and pathogens alike.

Chapter 5: Chemoreceptor Gene Loss and Pathogenicity in *Escherichia coli*

Publication (reproduced with permission from):

Chemoreceptor Gene Loss and Acquisition via Horizontal Gene Transfer in *Escherichia coli*. **Borziak K, Fleetwood AD, and Zhulin IB.** *Journal of Bacteriology* (2013). 195(16):3596-3602.²⁴² Copyright 2013, American Society for Microbiology.

Author Contributions:

K Borziak and AD Fleetwood contributed equally to the article. AD Fleetwood was responsible for elements of the paper related to pathogenicity types, physiology, and estimation of the divergence of the primarily extra-intestinal pathogen clade B2. This involved preliminary analyses, a comprehensive literature review, the sSNP molecular clock analysis, and investigation of a putative sucrose chemoreceptor gain event. Additionally, AD Fleetwood initially constructed and manually curated concatenated multiple sequence alignments of chemotaxis proteins. K Borziak obtained house-keeping gene and chemotaxis datasets, constructed phylogenetic trees, and analyzed the presence or absence of chemotaxis genes across all strains (including quality control for sequencing artifacts). K Borziak also investigated *aer* and *tsr* acquisition events. IB Zhulin initially observed the loss of *trg* and *tap* genes in *Escherichia coli* chemoreceptor sequences.

*Mobley Lab (University of Michigan) independently published experimental observations that *trg* and *tap* were missing from a urinary pathogenic *E. coli* strain and posited that these were lost due to a lack of selective pressure for their ligands.²⁴³ Upon discovery of this work, we consulted with the Mobley lab, who provided helpful experimental insight and suggestions.

Supplementary figures and tables from this publication are located in Appendix B and “Dataset B1.xlsx”. In addition to being reproduced and formatted for this dissertation, the publication has been revised with the addition of a section at the end of the results further addressing sSNP molecular clock analysis conducted by AD Fleetwood.

Abstract

Chemotaxis allows bacteria to more efficiently colonize optimal microhabitats within their larger environment. Chemotaxis in *Escherichia coli* is the best-studied model system and a large number of *E. coli* strains have been sequenced. The *Escherichia/Shigella* genus encompasses a great variety of commensal and pathogenic strains, but the role of chemotaxis in their association with the host remains poorly understood. Here we show that the core chemotaxis genes are lost in many, but not all, non-motile strains, but are well preserved in all motile strains. The genes encoding the Tar, Tsr and Aer receptors that mediate chemotaxis to a broad spectrum of chemical and physical cues are also nearly uniformly conserved in motile strains. In contrast, the clade of extra-intestinal pathogenic *E. coli* apparently underwent an ancestral loss of Trg and Tap chemoreceptors that sense sugars, dipeptides and pyrimidines. The broad range of time estimated for the loss of these genes (1-3 million years ago) corresponds to the appearance of the genus *Homo*.

Introduction

Escherichia coli are ubiquitous colonizers of the intestines of mammals and birds.⁹⁰ There are several highly adapted *E. coli* clones that have acquired virulence traits and cause a broad spectrum of disease including enteric/diarrheal disease, urinary tract infections (UTIs), and sepsis/meningitis.²⁴⁴ Depending on the site of infection, pathogenic strains are classified as intestinal (IPEC) and extraintestinal (ExPEC) pathogenic *E. coli*, and distinct pathotypes (based on clinical manifestation) are recognized within both categories. The most common ExPEC pathotypes include uropathogenic (UPEC), meningitis-associated (MNEC), and avian pathogenic (APEC) *E. coli* strains.^{244,245} Motility was shown to be important for the colonization of both commensal and pathogenic *E. coli*, as well as the pathogenesis of the latter:^{246,247} however, the exact role of motility and the underlying chemotaxis system in these processes remains poorly understood. Molecular machinery that controls chemotaxis in *E. coli* has been the subject of intensive investigation.^{31,131} Its components include chemoreceptors, also known as methyl-accepting chemotaxis proteins (MCPs), a histidine kinase CheA, an adaptor protein CheW, a methyltransferase CheR and a methylesterase CheB, as well as a response regulator CheY and its phosphatase CheZ. *E. coli* has five chemoreceptors. Tsr mediates attractant responses to serine and quorum autoinducer AI-2,^{248,249} as well as responses to oxygen, redox and oxidizable substrates.^{48,250} It was also recently shown to mediate taxis to 3,4-dihydroxymandelic acid, a metabolite of norepinephrine that is produced by human cells (Mike Manson, personal communication). Tar mediates attractant responses to aspartate and maltose^{249,251} and negative responses to metal ions.²⁵² Trg mediates attractant responses to ribose and galactose,²⁵³ while Tap does so for dipeptides and pyrimidines.^{45,46} Aer mediates responses to oxygen and energy taxis.^{47,48} The majority of the chemotaxis proteins are encoded in two adjacent operons, *mocha* (*motA*, *motB*, *cheA*, *cheW*) and *meche* (*tar*, *tap*, *cheR*, *cheB*, *cheY*, *cheZ*), whereas the remaining three chemoreceptors (Tsr, Trg, and Aer) are encoded elsewhere on the chromosome. On a large evolutionary scale, the chemotaxis system, which

appeared in a common ancestor of Bacteria, underwent drastic changes displaying a wide array of variations in component design.⁵⁰ Even the closest relatives of *E. coli* show substantial differences in the chemotaxis machinery. In *Salmonella enterica*, the majority of chemotaxis components are orthologous to those of *E. coli*, but it lacks Tap, and contains additional chemoreceptors and the second adaptor protein, CheV.²⁵⁴ However, the driving forces that shape the chemotaxis system on a small evolutionary scale remain unknown.

E. coli is the most sequenced bacterium to date and phylogenetic studies provided important insights into the processes of its genome evolution.²⁵⁵⁻²⁵⁷ *E. coli* strains are too closely related to each other to be resolved by classical 16S- and ribosomal protein-based phylogeny. Based on several other independent methods including multi-locus enzyme electrophoresis, multi-locus sequence typing, feature frequency profiles, and whole genome phylogeny *E. coli* strains are classified into several phylogenetic groups: A, B1, B2, D, E, and F.^{255,257-259} The phylogenetically defined *E. coli* clade²⁶⁰⁻²⁶² also includes *Shigella* clones that have been historically considered a separate genus due to distinct phenotypic features, such as loss of motility. Chemotaxis has been studied extensively using derivatives of a single *E. coli* strain, K-12 (the A group), and the functionality and conservation of the chemotaxis system has not been specifically studied in members of other *E. coli* groups. Several studies suggested the dispensability of both core and accessory chemotaxis components in *E. coli*. The core genome of *E. coli* contains nearly 2,000 genes.²⁵⁶ Interestingly, only a subset of the chemotaxis genes belongs to the core genome according to this study. Key components of the chemotaxis system, CheW and CheB as well as two major chemoreceptors, Tar and Tsr, are missing from this core set suggesting that chemotaxis might be a dispensable function in *E. coli*. Furthermore, several uropathogenic *E. coli* strains were shown to lack Trg and Tap receptors, and it was postulated that the gene loss was a result of a lack of selective pressure on sugar and peptide sensing receptors in the urinary tract, which is void of these substrates.²⁴³ Here, we analyzed the chemotaxis system of *E. coli* by comparing genomes of more than 200 strains that included

commensals and pathogens from all known phylotypes. We show that the chemotaxis system is well-preserved in *E. coli*, even among some strains that have lost motility and that the major evolutionary event was the loss of Trg and Tap receptors that occurred not only in some uropathogenic strains, but in the common ancestor of the B2 phylotype. We propose that losing the ability to sense sugars, peptides and nucleotides contributed to the emergence of extra-intestinal clones including pathogens.

Materials and Methods

Data sources and bioinformatics software. The following software packages were used in this study: HMMER v3.0,²⁷ Jalview,²⁵ MAFFT v6.847b,²⁶³ MEGA v4.0,²⁶⁴ PhyML v3.0,²⁶⁵ and BLAST+ v2.2.4+.²⁶⁶ All multiple sequence alignments were built in MAFFT with its I-INS-i algorithm. All maximum likelihood phylogenetic trees were built in PhyML with standard parameters and subtree pruning and regrafting topology search. Genomes, proteomes, and genome annotations of all distinct *Escherichia* and *Shigella* strains available in the NCBI *nr* database as of 12th January, 2012 were collected (219 genomes). Pathotype information was retrieved from primary literature and public databases including PATRIC and GOLD.^{267,268}

Construction of a phylogenetic tree for Escherichia. *Escherichia* phylogenetic tree was constructed using the *arcA*, *aroE*, *icd*, *mdh*, *mtlD*, *pgi*, and *rpoS* genes.²⁶⁹ The nucleotide sequence sets for each gene were aligned individually in MAFFT. The alignments were concatenated, and the resulting alignment was used to build a maximum likelihood tree in PhyML.

Identification of chemotaxis and accessory proteins in genomic data sets. Chemotaxis and accessory genes and proteins were retrieved from the genome of *E. coli* W3110 (model wild type for chemotaxis) and used as BLAST queries against the genome set. Protein and nucleotide searches were performed to ensure retrieval of missing and partial genes. Gene neighborhoods were extracted from NCBI genome feature files.

Multiple sequence alignment and phylogenetic analyses. The nucleotide and protein chemotaxis sequence sets (MotA, MotB, CheA, CheW, Tar, Tap, CheR, CheB, CheY, CheZ, Tsr, Trg, and Aer) were individually aligned by MAFFT. The alignments of the chemotaxis operons, *mocha* and *meche*, were concatenated and used to build a maximum likelihood tree in PhyML.

sSNP molecular clock calculation: All of the chemotaxis genes (except for *trg* and *tap*) and *recA* from clades B2 and A were individually aligned and concatenated to produce a gapless alignment. After removing sequences with errors, the final set consisted of 58 sequences (**Table S1** of Appendix B). The alignment spanned 4,360 codons. The equation (**Equation 1**) used to calculate time of divergence is: $(\text{number of sSNP sites}) / (\text{potential sSNP sites} \times \text{mutation rate} \times \text{generations per year} \times 2)$ Potential sSNP sites were determined using the parsimonious assumption that each codon has only one potential sSNP site. Generations per year were estimated at a range from 100 to 300 to allow for a broad estimation.²⁷⁰⁻²⁷³ The experimentally determined synonymous mutation rate of 1.4×10^{-10} was used.²⁷⁴

Results

Phylogenetic tree of Escherichia. We analyzed 219 (55 complete and 164 draft) genomes of *Escherichia* and *Shigella*. This set included genomes of *E. fergusonii* and *E. albertii*, to serve as outgroups in the phylogenetic analysis. In order to assign newly sequenced strains to the established phylogenetic groups, we have constructed a phylogenetic tree of all 219 strains in our dataset. Because relationships between such closely related strains cannot be resolved using traditional ribosomal trees, we built a maximum-likelihood tree from concatenated alignments of the *arcA*, *aroE*, *icd*, *mdh*, *mtlD*, *pgi*, and *rpoS* genes, as previously suggested²⁶⁹. The tree (**Figure S1** of Appendix B) is in good agreement with previously published data, including whole genome-based phylogeny.²⁵⁶ Detailed classification of all

Escherichia genomes based on pathotype and phylogenetic groups is shown in “**Dataset B1.xlsx**”.

Core chemotaxis genes. The presence and absence of eleven chemotaxis genes (*cheA*, *cheW*, *cheY*, *cheB*, *cheR*, *cheZ*, *tsr*, *tar*, *trg*, *tap* and *aer*) in all 219 genomes is shown as a bird-eye view in **Figure S2** of Appendix B. The picture looks like a mildly used shooting target: while concentric rings representing the presence of each of the chemotaxis proteins are well preserved, there are visible holes of different sizes showing the absence of particular genes. Many of the missing proteins can be found as pseudogenes resulting from single-nucleotide frameshifts. Sequencing errors (rate of 1% for some next-generation sequencing methodologies) appear to be the main source of missing proteins (e.g. *cheB* split as ECH7EC4401_1543 and ECH7EC4401_1544 in *E. coli* O157:H7 str. EC4401). Another common cause of missing genes in draft genomes is a split between different contigs (e.g. *cheA* split between ZP_04536326 and ZP_04536327 in *Escherichia* sp. 3_2_53FAA). An additional cause is erroneous gene calling (e.g. a complete *cheA* gene in *E. coli* str. K-12 substr. DH10 is missing). We have analyzed each and every potential mutation in all chemotaxis genes, assigning them to obvious sequencing, assembly, and annotation errors or potentially true mutations (“**Dataset B1.xlsx**”). Completely sequenced, closed genomes served as the main internal control. Distribution of chemotaxis genes in closed genomes only is shown in **Figure 24**.

To better discriminate between potential sequencing/assembly errors and true mutations, we analyzed the nature of mutations in *Shigella* genomes. *Shigella* are non-motile due to inactivation of their flagellar genes,^{275,276} therefore accumulation of mutations in their chemotaxis genes was expected. Indeed, 30% of *Shigella* strains had significant deletions and insertions in the *mocha/meche* operons (“**Dataset B1.xlsx**”). Deletions were present not only in draft, but also in complete genomes of *Shigella*, reducing the chance of these results being attributable to sequencing errors. Only 33% of *Shigella* strains contained complete sets of intact

chemotaxis genes. In a striking contrast, none of the *E. coli* strains has accumulated insertions or deletions in their core chemotaxis genes (*cheA*, *cheW*, *cheY*, *cheB*, *cheR*, and *cheZ*). Single frameshift mutations in these genes were identified only in nine *E. coli* genomes, all of which were in draft status and could be due to sequencing errors. All completely finished *E. coli* genomes had their core chemotaxis genes intact. No events of gene duplication or horizontal gene transfer have been found among core chemotaxis genes.

Chemoreceptor loss. In contrast to core chemotaxis genes, chemoreceptor loss was observed not only in *Shigella*, but also in some *E. coli* strains. In *Shigella*, all five chemoreceptors (Tar, Tsr, Trg, Tap, and Aer) have a nearly equal chance to be eliminated, whereas in *E. coli* chemoreceptor loss was strongly biased toward Trg and Tap (**Table 5**). Most strikingly, this loss was observed in specific phylotypes. All B2 group strains and the majority of F group strains underwent a deletion in the *tap* gene. The identical nature of the deletions (**Figure 25**) suggests that the event occurred prior to the B2 clade divergence. The majority (33 of 38) of B2 strains have also undergone a deletion in the *trg* gene. Similarly to the deletion of *tap*, the symmetrical nature of the *trg* deletion (**Figure 25**) suggests that the loss was an ancestral event. Another four B2 group strains possess an identical frameshift mutation within the *trg* gene. The symmetrical nature of this frameshift and its presence in a completely sequenced genome of the *E. coli* 536 strain (**Figure 25**) indicate that it is not a sequencing artifact. Thus, it appears that *trg* and *tap* deletions occurred in a common ancestor of a clade, which approximately corresponds to the B2 phylogroup. Using molecular clock calculations, we estimated a time period during which the ancestral chemoreceptor loss event occurred. We compared the number of synonymous mutations in the B2 clade in which the loss took place with the A clade that contains the chemotaxis wild-type strains K12. The B2 clade has overall and on average more sSNPs than the A clade, indicating a longer time period of divergence from respective common ancestors. Our estimates indicate that B2 diverged from ~1 to 3 million

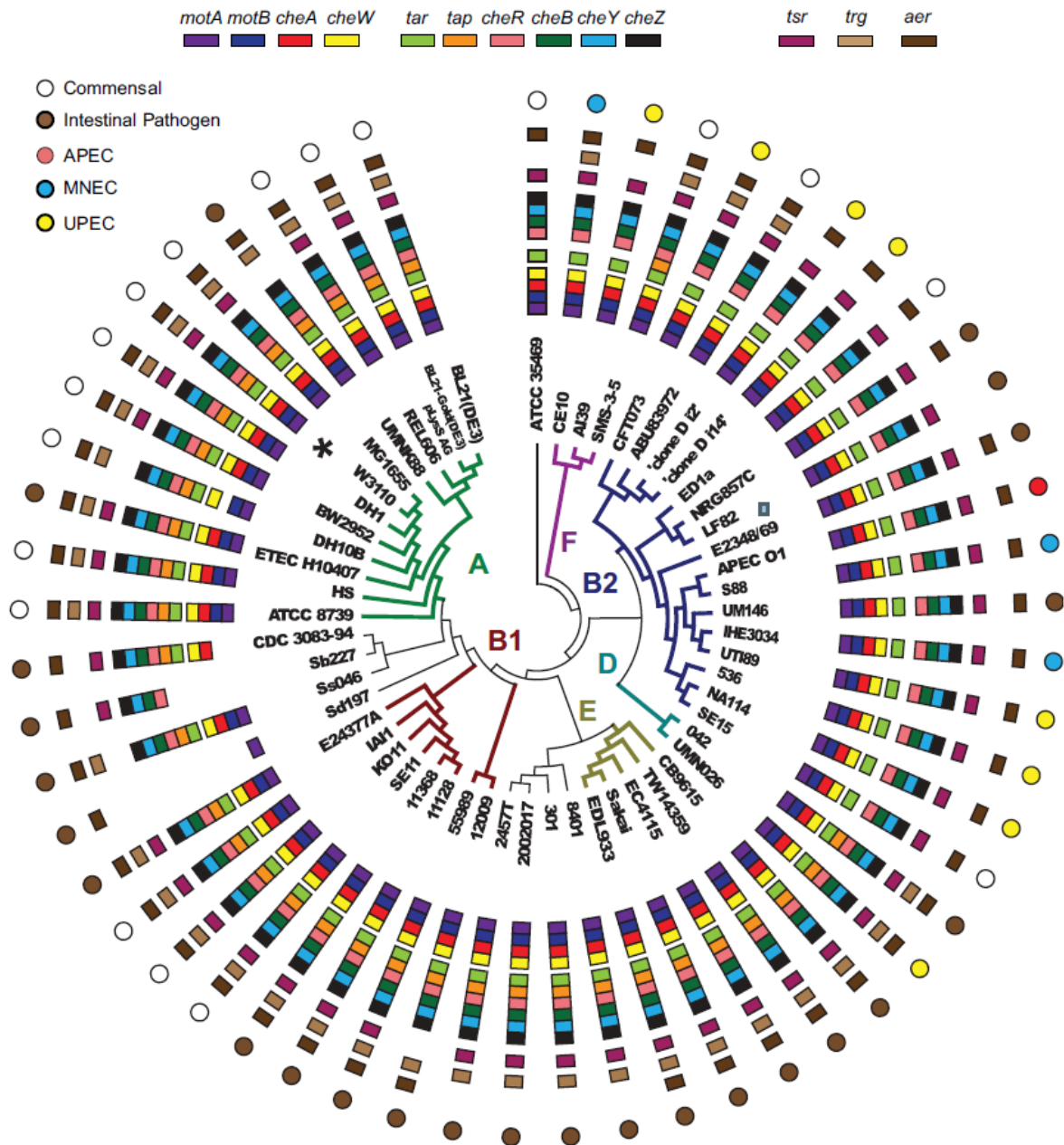


Figure 24. Presence of Chemotaxis Genes in Completely Sequenced *Escherichia/Shigella* genomes. Full strain names and properties are listed in “**Dataset B1.xlsx**”. Phylogenetic relationships are shown in the center. Branches are color coded according to previously established phylotypes. *E. coli* K-12 strain W3110 (model for chemotaxis) is marked with an asterisk. Outer ring denotes pathogenicity types of each strain.

Table 5. Loss of Chemoreceptor Genes in *E. coli* and *Shigella* Genomes.

Lost gene*	<i>E. coli</i> genomes		<i>Shigella</i> genomes	
	All (183)	Finished (46)	All (28)	Finished (8)
<i>tar</i>	0	0	12	2
<i>tsr</i>	4	2	4	1
<i>aer</i>	1	1	12	4
<i>trg</i>	34	16	7	3
<i>tap</i>	41	18	10	2

*Excluding detected sequencing/assembly/annotation errors (see **Dataset S1** for details)

years ago (Ma), whereas the A clade did so from ~0.4 to 1.2 Ma (300 and 100 generations per year respectively).

Chemoreceptor acquisition. While no chemoreceptor gene duplication was observed in any analyzed genome, we detected several receptor acquisition events (**Table 6**). All acquired MCPs were plasmid-borne. In *E. fergusonii* ECD227 an acquired MCP is 99% identical to the MCP from *Salmonella enterica* subsp. *enterica* serovar Kentucky str. CVM29188, which is also located on a plasmid. These plasmids are similar and were implicated in antimicrobial resistance in *Salmonella* and virulence in *E. fergusonii*.²⁷⁷ This chemoreceptor is significantly different from canonical *E. coli* MCPs in sequence, although it belongs to the same class 36H¹⁵³ and has the same predicted membrane topology. *E. coli* O157:H7 str. EC4024 acquired a MCP that was identified from its N-terminal portion (residues 1-350) located at a contig end. This fragment was 99% identical to an MCP from an *Enterobacter hormaechei* (GI: 334124148) and showed limited similarity to Trg (less than 30% identity). The MCP is found neighboring a sucrose metabolism gene cluster both on the plasmid and in the *Enterobacter* genomes, suggesting a possible role as a sucrose sensor. Finally, seven *E. coli* genomes were found to possess an *aer*-like MCP likely acquired from *Aeromonas caviae*, which is also known to cause

gastroenteritis.²⁷⁸ In six genomes, these MCPs are identical, suggesting a single recent acquisition event.

(Begin Additional Section: Text and Equation 1 only) *sSNP Molecular Clock Analysis*.

Both the authors and the Mobley Lab were interested in determining when the ancestral gene loss event occurred. The Mobley Lab had independently theorized (based on the lack of receptors in a urinary pathogenic strain) that lack of exposure to the ligands due to the new environment of the urinary tract relaxed selection pressure. However, since our study determined that the loss was unequivocally ancestral and resulted in the generation of a variety of extra-intestinal pathogens (not just urinary pathogens), we concluded that the loss may have initially decreased competitiveness for resources in *E. coli*'s preferred habitat (the gastrointestinal tract), which forced these strains to adapt to either more efficiently colonize the intestines or move to new niches.

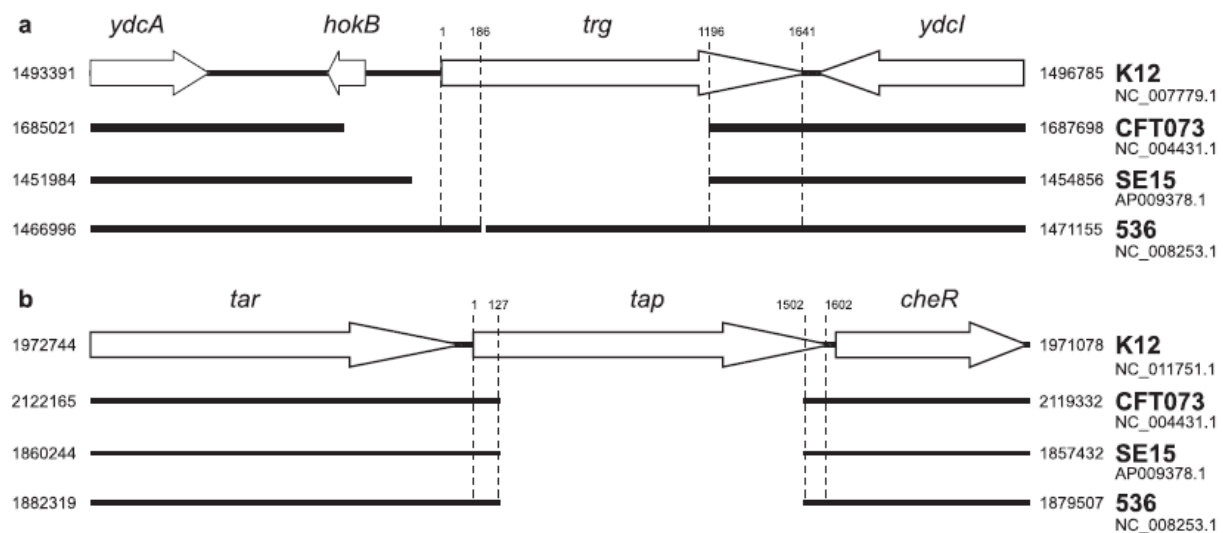


Figure 25. Deletions in *tap* and *trg* Genes in B2 Group Strains. Gene neighborhoods in representative genomes are shown. Full strain names and genomic locations of deletions are listed in “**Dataset B1.xlsx**”.

Table 6. Horizontally Transferred Chemoreceptor Genes in *Escherichia* Genomes

Genome	Acquired gene		Closest BLAST hit	
	Name GI	Sequence Identity with <i>E. coli</i> K-12 homolog	Organism GI	Sequence Identity
<i>E. fergusonii</i> ECD227	Tsr (MCP I) 424819104	37%	<i>S. enterica</i> 194447140	99%
<i>E. coli</i> O157:H7 str. EC4024	Trg (MCP III) 195941089	29%	<i>E. hormaechei</i> 334124148	99%
<i>E. coli</i> 101-1	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	99%
<i>E. coli</i> E1520	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%
<i>E. coli</i> G58-1	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%
<i>E. coli</i> MS 84-1	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%
<i>E. coli</i> MS 85-1	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%
<i>E. coli</i> MS 124-1	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%
<i>E. coli</i> TA007	AER (MCPV) 19443928	33%	<i>A. caviae</i> 51470604	100%

In order to estimate a timeline for this event, we needed to establish a comparison, as estimations are only relative. Thus, we chose to compare the B2 group with the ancestral gene loss to the A group, which are comprised primarily of human commensal strains. Our justification for this comparison is that the A group appears to be the evolutionarily youngest (fewest number of synonymous mutations), implying that it has had the least amount of time to diverge since it branched off from the rest of *E. coli*. Since this group is primarily composed of commensal organisms, this looks to be the most recent branch of human specialist *E. coli*, providing a perfect contrast to the most highly specialized pathogenic group, B2.

The basic theory behind the molecular clock calculation hinges on the concept of sSNPs (synonymous mutations/nucleotide polymorphisms).²⁷⁰⁻²⁷³ Synonymous mutations are nucleotide changes that change the codon but not the protein for which it codes (most amino

$$\frac{(\# \text{ of synonymous Single-Nucleotide Polymorphisms [sSNP] sites})}{(\# \text{ of potential sSNP sites} \times \text{mutation rate} \times \text{generations per year} \times 2)}$$

Where: # of potential sSNP sites (4,360)
 mutation rate = 1.4×10^{-10}
 generations per year = 100, 200, or 300

Equation 1. sSNP Molecular Clock Analysis for B2 and A Group Diversification

acids are encoded by 4 codons, though this is variable). These are important because they do not alter the structure or function of the protein (which is a valid assumption, though certain rare instances of nucleotide changes can alter translation rates or other properties of the protein that result in altered function). Since synonymous mutations aren't dramatically altering function, they are not changing evolutionary pressure on the gene, and they will naturally accrue over time at a somewhat predictable rate. This rate is based on the error rate of DNA polymerases, experimentally observed in *E. coli*.²⁷⁴

Turning to the equation itself, the key data on which the quality of the molecular clock estimation relies is the number of sSNPs. In order to detect these events, all of the chemotaxis genes for each group were separately aligned (nucleotide sequences, not amino acid sequences) using MAFFT. Thus, chemoreceptor orthologs were aligned together, CheA genes were aligned together, and then these individual high quality alignments were concatenated. When genes were missing or divergent (i.e. the missing *trg* and *tap* chemoreceptors in B2 group), these were removed from the alignment so that the comparison would not be biased by genes that were no longer under the same evolutionary pressure (the pseudogene remnant of these genes would have accrued mutations at greatly accelerated rates since there was no longer a protein to function). The final datasets consisted of two concatenated nucleotide

alignments of 13,890 nucleotides (4,630 codons) each (one for B2 and one for A). The imbalance in the number of sequences is controlled for by normalizing the number of sSNPs by the number of potential sSNP sites (see **Equation 1**), which simplified is the number of codons multiplied by the number of sequences within the alignment.

Other factors in the analysis include factoring in the number of generations per year (estimated conservatively at 100 and less strictly at 300), providing the range for the estimation. This is important, as various generation times have been observed under an assortment of environmental and experimental conditions. The number of mutations increases as the generational time decreases, creating a broad estimation as opposed to a pinpoint result (which would be specious). Finally, the fact that bacterial doubling produces two new strands of each gene, and thus two chances for each mutation, is also factored into the equation. Our resulting estimate was ~1-3 million years ago. **(End of Additional Section)**

Discussion

Despite a relatively short timeline of divergence, the chemotaxis system in the genus *Escherichia* has undergone substantial changes. First, the loss of the entire chemotaxis function manifested as severe mutations in core chemotaxis genes was observed. This event was unambiguously detected only in non-motile, intracellular *Shigella*. All *E. coli* genomes contain intact core chemotaxis genes indicating that chemotaxis is critical for motile strains. On the other hand, not all *Shigella* lost their chemotaxis genes. For example, in the *S. flexneri* K-671 the entire chemotaxis system appears to be intact, whereas flagella are absent due to mutations in the *flhDC* flagellar master operon.²⁷⁹ Several *Shigella* strains retain intact *mocha* and *meche* operons. Thus, the chemosensory apparatus in these strains might be used for other functions. This is a common trend in the evolution of the chemotaxis system on a larger evolutionary scale: it was co-opted to control such processes as gene expression in many bacterial

species.^{50,280} Second, we detected changes in the chemoreceptor repertoire caused by gene loss and, to a lesser extent, by horizontal gene transfer, but not gene duplication. The major chemoreceptors Tar and Tsr are well preserved in *E. coli*. This is consistent with their roles as modulators of important behaviors that in addition to sensing various attractants and repellents include energy taxis,⁴⁸ thermotaxis,²⁸¹ and pH taxis.²⁸² Tar and Tsr are equally important for commensal and pathogenic strains. These chemoreceptors are also necessary and sufficient for chemotaxis toward urine in the pathogenic *E. coli* strain CFT073.²⁸³ Although the aerotaxis receptor Aer has been categorized as a minor receptor according to its low abundance in the cell,²⁸⁴ it is also well preserved in *E. coli*, likely due to its role in energy taxis and thermotaxis. Consequently, we propose to refer to Aer as a major chemoreceptor, in addition to Tar and Tsr.

We have found evidence for at least three independent events of new chemoreceptor acquisitions by *E. coli* strains. A Trg-like chemoreceptor was found to be encoded in a sucrose metabolism gene cluster. Both gene order conservation for this receptor (together with fructokinase) in *Enterobacteriaceae* plasmids and the known role for Trg to mediate chemotaxis to ribose and galactose suggest that it might sense sucrose. Sucrose and fructose metabolism gene clusters have been reported in several *E. coli* extra-intestinal strains.^{285,286} Another interesting case is an additional Aer-like chemoreceptor, which is present in several *E. coli* strains, but appears to be a result of a single acquisition event. Multiple copies of Aer are not uncommon among *Gammaproteobacteria*. For example, they are present in such pathogens as *Vibrio cholerae*²⁸⁷ and *Pseudomonas aeruginosa*.²⁰³

Unambiguously, loss can be established only for Trg and Tap, where large deletions were identified in corresponding genes in many *E. coli* genomes. The overwhelming majority of these strains belong to the B2 clade, which contains major extra-intestinal pathogens. The deletions occurred in the same chromosomal position in all B2 strains strongly suggesting a single ancestral event. This instance of gene loss from the chemotaxis system was a significant evolutionary event that had the potential to affect and shape the phenotypic behavior (including

emergence as a pathogen) that differentiates the B2 clade from those that still possess the canonical chemoreceptor genes. While there are several reported differences between B2 and other clades, extra-intestinal pathogenicity is the most striking, and its major impact on human well-being and the substantial associated costs of healthcare merits further investigation.

As to the evolutionary context of this gene loss, it does not appear to be a result of relaxed selective pressure on sensors to sugars and dipeptides that are exceedingly rare in urine from individuals with healthy kidneys. Genomes that lost *trg* and *tap* contain intact genes coding for ribose, galactose/glucose, and dipeptide periplasmic-binding proteins that mediate the sensing of these compounds through Trg and Tap. This suggests continuing importance of these metabolites (and therefore significant exposure to them), which is not in line with a selection driven loss due to minimal or non-exposure. Furthermore, B2 strains colonize the intestines very effectively²⁸⁸ and function as commensals until they are outside of the intestinal tract, so they are not exclusively under selection pressure from the urinary environment. Finally, once they have exited the intestinal tract, some extra-intestinal B2 strains are not found in the urinary tract but preferentially migrate elsewhere (hence MNEC and APEC strains). Thus, it is likely that the ancestral loss of *trg* and *tap* predisposed gut-inhabiting strains to seek other niches to occupy (either in the form of a new adaptive strategy in the gut or colonization elsewhere).

The molecular clock analysis of the chemotaxis system of the B2 strains suggests that they branched off fairly early (relatively speaking), which provides the ancestral receptor loss a long period of divergence over which its effects would be present. Even with as broad an estimation as ~1 to 3 Ma, this places the divergence of the B2 clade in the ballpark of the estimated appearance of the genus *Homo* (2.3-2.4 Ma)²⁸⁹ and provides an intriguing temporal link to human specialization and pathogenicity. Taken as a whole, we hope that this message encourages further discussion on the evolutionary history of *Escherichia coli* and also highlights

a potentially significant and novel method through which chemotaxis receptors might influence new pathogen emergence.

Acknowledgements

We thank Harry L. T. Mobley for discussion and helpful suggestions and Michael D. Manson for communicating results prior to publication and helpful suggestions.

This work was supported by the National Institute of Health grant GM072295.

Chapter 6: Conclusions

Chemotaxis is a mature field; it has been studied for almost fifty years and has been highly characterized from every experimental angle. Both the signal transduction and microbiology communities have made great strides recently, but more can be achieved by more comprehensive study of chemoreceptors. The molecular lessons of signal transduction work may ultimately translate into therapeutic intervention for chemotactic pathogens, while the wealth of experimental observations across different chemotaxis systems may likewise better refine our understanding of how biological signals are transduced in bacteria and human receptors alike. Confronting microbial pathogens is a dire necessity in light of rising global antibiotic resistance, and better insight into our own receptors will further rationalize the field of drug discovery and design (improving safety and efficacy). Already, comparative genomics studies have rapidly expanded and accelerated our understanding of chemoreceptors and chemotaxis at a systems level.^{50,153}

This dissertation work has contributed to the study of chemoreceptors by using the theory and insight gained from comparative techniques. Specifically, we first refined molecular interactions in *Thermotoga maritima* by comparing the evolutionary histories of interacting proteins in conflicting Chemoreceptor:CheA:CheW co-crystal structures. Our results provide genomic evidence that both CheW and CheA compete for the same region of the receptor tip, and highlight which elements of each crystal structure agree with evolutionary information. Next, we assisted in the characterization of a novel galactose-sensing chemoreceptor from *Campylobacter jejuni* and demonstrated that it was a recent innovation that was the result of duplication and domain swapping of two closely related *Campylobacter* receptors. Third, we pioneered a novel genomic approach for disentangling multiple chemoreceptors and multiple chemotaxis, showing for the first time a comprehensive model for the relationships between chemotaxis systems in *Pseudomonas aeruginosa* using phylogenetic and comparative analyses. Finally, we investigated evolution at the species level in *Escherichia coli* and connected ancestral

loss of chemoreceptor genes with extra-intestinal pathogenicity. The significance of this finding may not be limited to *E. coli*, but serves as an example of the potential for chemoreceptor repertoires to influence commensal to pathogen transitions in motile bacteria.

There are three major open questions facing the chemoreceptor community. First, how can we use bioinformatics methods to computationally predict the sensory specificity of a chemoreceptor? Without this information, most chemoreceptors are biologically irrelevant. Second, what is the function of soluble chemoreceptors in the context of the chemotaxis system, and where do they fit with respect to the prototypical transmembrane system? Soluble receptors have been largely disregarded until the last few years, and were all together written off by the signal transduction community (until our work in **Chapter 2**.) Now there are several examples that have been partially characterized, opening the doors for large scale comparative work. Finally, how do multiple chemotaxis systems co-exist within an organism, and how do these systems regulate one another so that conflicting motility signals do not jeopardize the survival of the organisms (especially with multiple flagellar systems)? Already, we have seen chemoreceptors located in one chemotaxis gene cluster but localized to another,¹¹¹ and this may only be the tip of the iceberg as to the complex interchange and cross-talk that may be possible between systems. Conversely, if there is no cross-talk, the mechanisms of system specificity may reveal untold additional layers of regulation that will have general implications for localizing cells to the membrane, partitioning systems, and the overall molecular logistics of microbial organisms. We conclude with possible solutions for future comparative genomics work on chemoreceptor sequence, structure, and function.

With respect to predicting chemosensory capabilities, current domain models face shortcomings in both sensitivity and specificity. In the first case, our lab has had success using profile-profile comparisons (as opposed to searching sequences with a profile), which greatly increases the sensitivity of protein domain identification (see **Figure 26** for an example from

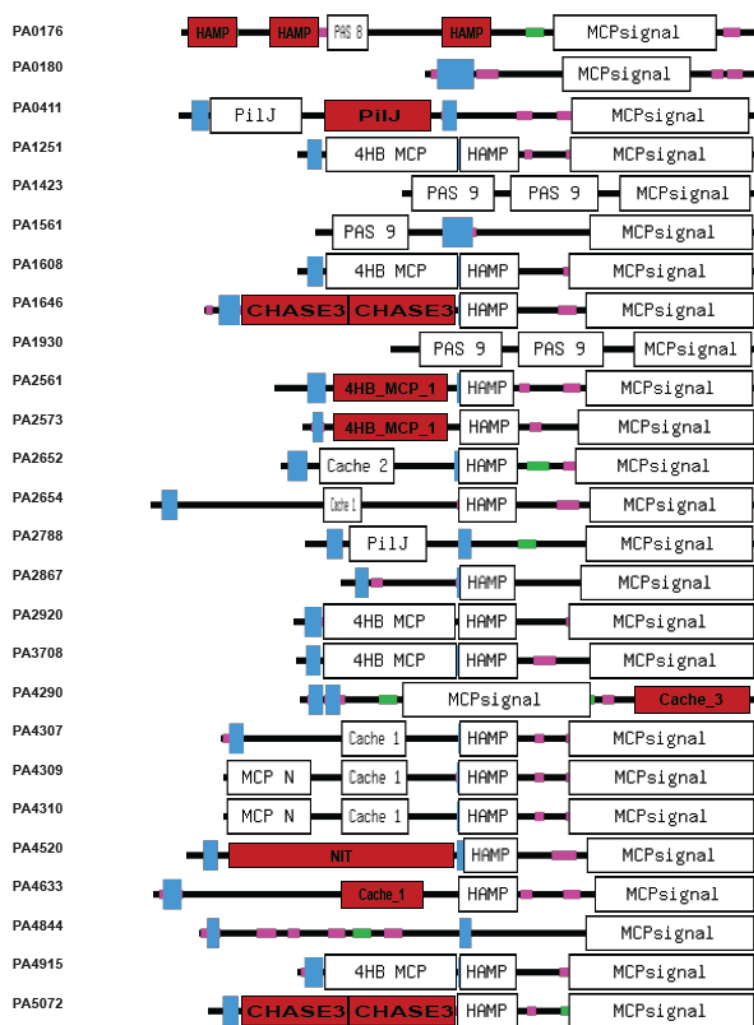


Figure 26. Filling in the Gaps of Chemoreceptor Sensory Capabilities. Of the 26 chemoreceptors in *P. aeruginosa* PAO1, almost half (12/26) have significant portions of their sequence not covered by current domain models using Pfam profile-sequence searches. By using more sensitive profile-profile searching methods (HHpred, Max Planck Institute), the vast majority of these regions can be linked to known sensory domain models (colored in Red). Already, advances in comparative methods are capable of filling in these gaps that riddle protein space. These models may one day be refined to the point where specificity may be heavily suspected (e.g: Cache: Amino Acids, 4HB: Pyrimidine Nucleotides, etc.). The NIT domain is a perfect example of one of the few domains that currently deliver this level of information (see **PA4520**, red domain fifth from the bottom), as we now know through our analysis that this chemoreceptor most likely senses nitrites and nitrates. Nitrite and nitrate sensing has been reported in *P. aeruginosa*, but no chemoreceptor was ever implicated before this type of analysis.

work on *P. aeruginosa*). As for the second problem, we contend that we have the tools and resources available to further sub-divide current domain models until they provide concrete functional predictions. Sequence alone may not initially be enough, but the addition of experimental binding data (i.e. crystal structures or mutational studies) from a few key examples will pave the way for progress in this respect. Accomplishing this feat would have profound signal-transduction-wide implications for drug design, drug delivery, bio-remediation, bioenergy, environmental carbon cycling, and a host of other applications.

However, before this can be done *in silico* and *en masse*, a proof of concept will be necessary, and our work with the Korolik Lab in *Campylobacter jejuni* could provide the perfect set of receptors with which to do so. In addition to having five chemoreceptors with related sensory domains (3 of which have ligands identified), one of these, the *C. jejuni* aspartate receptor CcaA, is a very interesting evolutionary case. Aspartate is a critical metabolite and building block, as it is the precursor for a wide variety of compounds including other amino acid biosynthesis, so acquiring specificity for aspartate is not an insignificant biological accomplishment. As mentioned in earlier sections, aspartate receptors are prevalent in model organisms (like Tar in *E. coli* and *Salmonella enterica*), yet CcaA and Tar have done so with different ligand binding domains, which have arisen from completely distinct structural lineages. The aspartate receptor CcaA appears to be widely disseminated through *Epsilonproteobacteria*, and may be related to characterized *H. pylori* chemoreceptors, whereas Tar is widely distributed throughout enteric *Gammaproteobacteria*. An analysis of these two groups of chemoreceptors may lead to a better understanding of how different sensory mechanisms can converge to sense the same signal, allowing for the eventual creation of highly specific, curated models.

As for the function of soluble chemoreceptors, the answer may be closer than one might expect. Cytoplasmic chemoreceptors AerC, BdlA, hemAT, and McpB all sense metabolic/redox signals.^{57,231,290,291} These chemoreceptors are found across diverse phyla (*Proteobacteria* and

Firmicutes), and their sensory domains are wholly unrelated (double PAS vs. Protoglobin). In our work on *T. maritima* from **Chapter 2**, we also worked with a soluble chemoreceptor from a deeply branching phylum, but this receptor (TM0014) had no predicted function and not enough unknown sequence to house a hidden sensory domain. Dr. Zhulin noticed a trend in this receptor's gene neighborhood: a close-association with a predicted redox-sensing beta-lactamase. Bipartite (split into two genes) chemoreceptors are not without precedent, as the signaling domain can associate with the sensory domain when they are co-expressed.⁶³ We investigated this putative beta-lactamase, and our preliminary results indicate that this gene neighborhood connection is present for the vast majority of easily identifiable TM0014 orthologs. The beta-lactamase from also appears to match known structures of di-iron redox sensors. Therefore, there appears to be a universal need for using soluble chemoreceptors for internal metabolic sensing (see **Figure 27**). These chemoreceptors are not bound to the membrane, and may be free to shuttle between systems and couple system activity, serving as global regulators. These proteins offer a compelling evolutionary story (convergent evolution), made even more intriguing by protoglobin's further ancestral link to hemoglobin.²⁹⁰ On the applied side, these potential master regulators may also provide a novel class of antibiotic targets.

Finally, returning to the last question facing chemotaxis, the multitude of chemoreceptors and multiple chemotaxis systems, the next clear step is to apply our comparative methodology to other organisms. The pathogen *Vibrio cholerae* (3 chemotaxis systems and 45 chemoreceptors) will offer an increased challenge with more chemoreceptors; however, if the trend from *P. aeruginosa* holds, there will be one major system with an associated class of receptors and two highly specialized systems. Both a future *Vibrio* study and the current *Pseudomonas* study can be combined and extended further to include increasingly deeper taxonomic levels (e.g. *Vibrionales/Pseudomonadales*, *Gammaproteobacteria*, *Proteobacteria*). In doing so, we can identify orthologs, divergent homologs, horizontal acquisitions, and novel

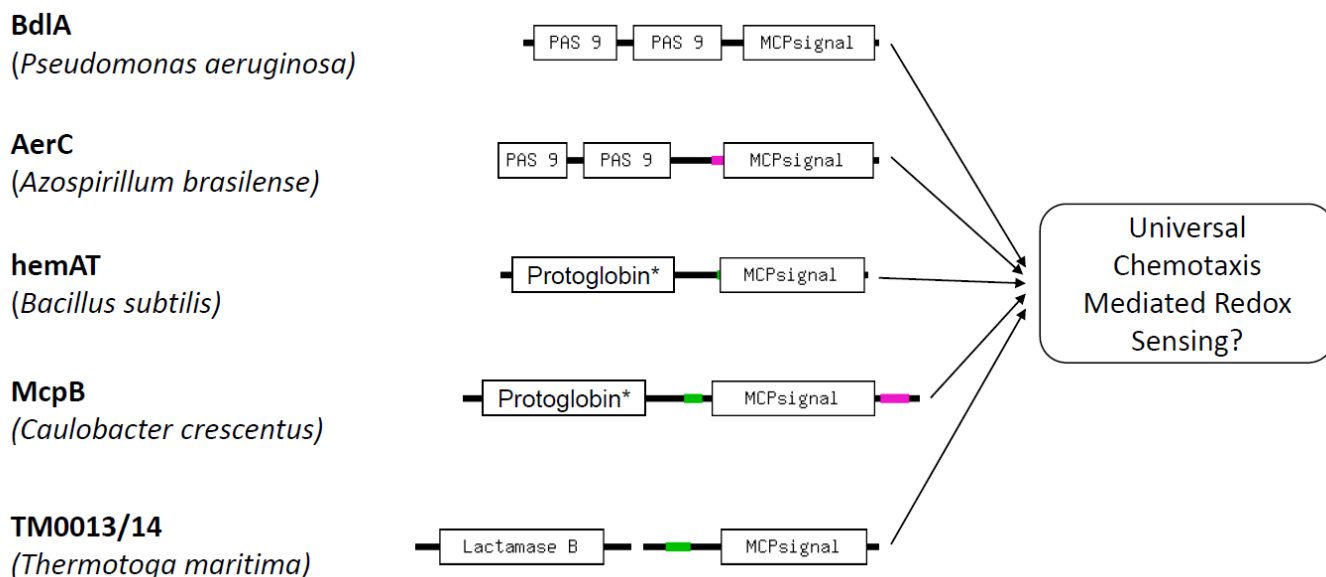


Figure 27. Convergent Evolution of Soluble Metabolism-Sensing Chemoreceptors?

Representatives of characterized or partially characterized soluble chemoreceptors. Domain architecture obtained from MiST 2.2 database. (*Protoglobin domains not present in MiST but are significant domain hits with experimental confirmation). BdlA, AerC, and hemAT are all experimentally confirmed redox sensors. McpB is a homolog of hemAT. Both BdlA and McpB have experimental evidence of proteolytic post-processing. TM0014 has been co-crystallized, but no sensory domain is predicted and the uncovered portion of the sequence is likely too short for a full domain. TM0013 is a divergent beta-lactamase domain containing protein whose orthologs are co-encoded in adjacent to TM0014's orthologs in phylum *Thermotogae*. This correlation holds for a wide diversity of bacterial phyla, but more robust analysis is needed. (Fleetwood and Zhulin, unpublished data)

innovations. This type of study would approach the scale of a chemoreceptor pan-genome, which would elucidate the importance of chemoreceptors in the context of different microbial lifestyles (plant, animal, and human pathogens and non-pathogens alike). This type of comparative analysis may also yield core ancestral chemoreceptors that could serve as broad

spectrum antibiotic targets, as well as individual chemoreceptors that could serve as species and even strain specific targets or biomarkers.

Comparative genomics of microbial chemoreceptors stands to greatly expand our understanding of their sequence, structure and function. Chemoreceptors serve as models for many of our own receptors (i.e. GPCRs), and they also show evidence of general features of protein evolution. Mechanistic studies of chemotaxis system formation have improved our understanding of the system, but we have yet to see many connections to the lifestyle of the organism. Conversely, lifestyle focused works have shown the potential impact and importance of chemoreceptors for health and the environment, but without deeper molecular and mechanistic understanding there are few options for application and intervention.

Chemoreceptors are under-appreciated virulence factors, as altered chemoreceptor suites can serve as corrupted GPS systems, directing commensals from favored niches to sites of novel or enhanced pathogenicity. As such, they are a major untapped sources of a novel antibiotic class in a world plagued by rising antibiotic resistance. By bridging the chemoreceptor signal transduction/microbiology gap, comparative genomics may usher in a new era of expanded significance to the field of chemotaxis, and may further the field's already storied contributions to the biological sciences as a whole.

List of Reference

- 1 Lander, E. *et al.* (International Human Genome Sequencing Consortium). Initial sequencing and analysis of the human genome. *Nature* **409**, 860-921 (2001).
- 2 Venter, J. *et al.* The sequence of the human genome. *Science* **291**, 1304-1351 (2001).
- 3 Creighton, T. Proteins: Structures and Molecular Properties (2nd ed.). *W.H. Freeman and Company* (1993).
- 4 IUPAC. (International Union of Pure and Applied Chemistry) Compendium of Chemical Terminology Gold Book v2.3.3. (2014).
- 5 Wheelan, S. J., Marchler-Bauer, A. & Bryant, S. H. Domain size distributions can predict domain boundaries. *Bioinformatics* **16**, 613-618 (2000).
- 6 Brocchieri L, K. S. Protein length in eukaryotic and prokaryotic proteomes. *Nucleic Acids Research* **33**, 3390-3400 (2005).
- 7 Koonin, E. Orthologs, Paralog, and Evolutionary Genomics. *Annu Rev Genet* **39**, 309-338 (2005).
- 8 Li WH, Y. J., Gu X. Expression divergence between duplicate genes. *Trends Genet* **21**, 602-607 (2005).
- 9 Woese CR, F. G. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc Natl Acad Sci U S A* **74**, 5088-5090 (1977).
- 10 Notredame C, H. D., Heninga J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol* **302**, 205-217 (2000).
- 11 Edgar, R. MUSCLE: Multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**, 1792-1797 (2004).
- 12 Wuichet K, A. R., Zhulin IB. Comparative Genomic and Protein Sequence Analyses of a Complex System Controlling Bacterial Chemotaxis. *Methods Enzymol* **422**, 1-31 (2007).
- 13 Katoh K, S. D. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**, 9 (2013).
- 14 Ulrich, L. E. & Zhulin, I. B. Four-helix bundle: A ubiquitous sensory module in prokaryotic signal transduction. *Bioinformatics* **21**, iii45-48 (2005).
- 15 Larkin MA, B. G., Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, Thompson JD, Gibson TJ, Higgins DG. Clustal W and Clustal X version 2.0. *Bioinformatics* **23**, 2947-2948 (2007).
- 16 Nei, M. Phylogenetic analysis in molecular evolutionary genetics. *Annu Rev Genet* **30**, 371-403 (1996).
- 17 Jones DT, T. W., Thornton JM. The rapid generation of mutation data matrices from protein sequences. *Computer Applications in the Biosciences* **8**, 275-282 (1992).
- 18 Felsenstein, J. Evolutionary Trees from DNA Sequences: A Maximum Likelihood Approach. *J Mol Evol* **17**, 368-376 (1981).
- 19 Bilwes, A. M., Alex, L. A., Crane, B. R. & Simon, M. I. Structure of CheA, a signal-transducing histidine kinase. *Cell* **96**, 131-141 (1999).
- 20 Bhatnagar, J., Borbat, P, Pollard, A.M. Bilwes, A.M., Freed, J.R., Crane, B.R. Structure of the ternary complex formed by a chemotaxis receptor signaling domain, the CheA histidine kinase and the coupling protein CheW as determined by pulsed dipolar ESR spectroscopy. *Biochemistry* **49**, 3824-3841 (2010).
- 21 Ulrich LE, Z. I. The MiST2 database: a comprehensive genomics resource on microbial signal transduction. *Nucleic Acids Research*, doi:doi:10.1093/nar/gkp940 (2010).
- 22 Berman HM, W. J., Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. *Nucleic Acids Research* **28**, 235-242 (2000).
- 23 Punta M, C. P., Everhardt RY, Mistry J, Tate J, Boursnell C, Pang N, Forslund K, Ceric G, Clements J, Heger A, Holm L, Sonnhammer ELL, Eddy SR, Bateman A, Finn RD. The Pfam protein families database. *Nucleic Acids Research*, 12 (2012).

- 24 Cserzo M, W. E., Simon I, von Heijne G, Elofsson A. Prediction of transmembrane alpha-helices in procariotic membrane proteins: the Dense Alignment Surface method. *Prot Eng* **10**, 673-676 (1997).
- 25 Waterhouse AM, P. J., Martin DMA, Clamp M, Barton GJ. Jalview version 2: A Multiple Sequence Alignment and Analysis Workbench. *Bioinformatics* **25**, 1189-1191 (2009).
- 26 Tamura K, S. G., Peterson D, Filipski A, Kumar S. MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular Biology and Evolution* **30**, 2725-2729 (2013).
- 27 Eddy, S. Profile hidden Markov models. *Bioinformatics* **14**, 755-763 (1998).
- 28 Eddy, S. A new generation of homology search tools based on probabilistic inference. *Genome Inform* **23**, 205-211 (2009).
- 29 Eddy, S. Accelerated Profile HMM Searches. *PLoS Comput Biol* **7**, e1002195 (2011).
- 30 Johnson LS, E. S., Portugaly E. Hidden Markov model speed heuristic and iterative HMM search procedure. *BMC Bioinformatics* **11** (2010).
- 31 Wadhams, G. H. & Armitage, J. P. Making sense of it all: bacterial chemotaxis. *Nat Rev Mol Cell Biol* **5**, 1024-1037 (2004).
- 32 Pazos, F. & Valencia, A. Protein co-evolution, co-adaptation and interactions. *The EMBO journal* **27**, 2648-2655, doi:10.1038/emboj.2008.189 (2008).
- 33 Rao, V. S., Srinivas, K., Sujini, G. N. & Kumar, G. N. S. Protein-Protein Interaction Detection: Methods and Analysis. *International Journal of Proteomics* **2014**, doi:10.1155/2014/147648 (2014).
- 34 Piasta, K. N., Ullmann, C. J., Slivka, P. F. & Falke, J. J. Elucidating Receptor-CheA Kinase Contacts in the Membrane-Bound Chemosensory Array of Bacteria using Disulfide Mapping and TAM-IDS. *Biochemistry* (2013).
- 35 Karplus, M. & Petsko, G. A. Molecular dynamics simulations in biology. *Nature* **347**, 631-639 (1990).
- 36 Brooks, B. R. et al. CHARMM: the biomolecular simulation program. *J Comput Chem* **30**, 1545-1614 (2009).
- 37 Draczkowski P, M. D., Jozwiak K. Isothermal titration calorimetry in membrane protein research. *Journal of Pharmaceutical and Biomedical Analysis* **87**, 313-325 (2014).
- 38 Patching, S. G. Surface plasmon resonance spectroscopy for characterisation of membrane protein-ligand interactions and its potential for drug discovery. *Biochimica et Biophysica Acta* **1838**, 43-55 (2014).
- 39 Adler, J. Chemotaxis in Bacteria. *Science* **153**, 708-716 (1966).
- 40 Hazelbauer, G. L. Bacterial Chemotaxis: The Early Years of Molecular Studies. *Annu Rev Microbiol* **66**, 285-303 (2012).
- 41 Adler, J. Chemoreceptors in Bacteria. *Science* **166**, 1588-1597 (1969).
- 42 Clarke S, K. J. D. Membrane receptors for aspartate and serine in bacterial chemotaxis. *J Biol Chem* **254**, 9695-9702 (1979).
- 43 Kondoh H, B. C., Adler J. Identification of a Methyl-Accepting Chemotaxis Protein for the Ribose and Galactose Chemoreceptors of Escherichia coli. *Proc Natl Acad Sci U S A* **76**, 260-264 (1979).
- 44 Wang EA, M. K., Clegg DO, Koshland Jr. DE. Tandem duplication and multiple functions of a receptor gene in bacterial chemotaxis. *J Biol Chem* **257**, 4673-4676 (1982).
- 45 Manson MD, B. V., Brade G, Higgins CF. Peptide chemotaxis in E. coli involves the Tap signal transducer and the dipeptide permease. *Nature* **321**, 253-256 (1986).
- 46 Liu X, P. R. Chemotaxis of Escherichia coli to pyrimidines: a new role for the signal transducer tap. *J Bacteriol* **190**, 972-979 (2008).
- 47 Bibikov SI, B. R., Rudd KE, Parkinson JS. A signal transducer for aerotaxis in Escherichia coli. *J Bacteriol* **179**, 4075 (1997).
- 48 Rebbapragada A, J. M., Harding GP, Zuccarelli AJ, Fletcher HM, Zhulin IB, Taylor BL. The Aer protein and the serine chemoreceptor Tsr independently sense intracellular

- energy levels and transduce oxygen, redox, and energy signals for *Escherichia coli* behavior. *Proc Natl Acad Sci U S A* **94**, 10541-10546 (1997).
- 49 Pruitt K, B. G., Tatusova T, Maglott D. The Reference Sequence (RefSeq) Database. *The NCBI Handbook* **Ch. 18** (2012).
- 50 Wuichet K, Z. I. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal* **3** (2010).
- 51 Hazelbauer, G. L. Adaptation by target remodelling *Nature* **484**, 173-175 (2014).
- 52 Kim, K. K., Yokota, H., and Kim, S.H. Four-helical-bundle structure of the cytoplasmic domain of a serine chemotaxis receptor. *Nature* **400**, 787-792 (1999).
- 53 Van Der Werf P, K. J. D. Identification of a gamma-glutamyl methyl ester in bacterial membrane protein involved in chemotaxis. *J Biol Chem* **252**, 2793-2795 (1977).
- 54 Ahlgren JA, O. G. Methyl esterification of glutamic acid residues of methyl-accepting chemotaxis proteins in *Bacillus subtilis*. *Biochem J* **213**, 759-763 (1983).
- 55 Parkinson, J. S. in *Annual Review of Microbiology*, Vol 64, 2010 Vol. 64 *Annual Review of Microbiology* (eds S. Gottesman & C. S. Harwood) 101-122 (2010).
- 56 Wadhams GH, M. A., Porter SL, Maddock JR, Mantotta JC, King HM, Armitage JP. TlpC, a novel chemotaxis protein in *Rhodobacter sphaeroides*, localizes to a discrete region in the cytoplasm. *Mol Microbiol* **46**, 1211-1221 (2002).
- 57 Xie, Z., Ulrich, L. E., Zhulin, I. B. & Alexandre, G. PAS domain containing chemoreceptor couples dynamic changes in metabolism with chemotaxis. *Proc Natl Acad Sci U S A* **107**, 2235-2240 (2010).
- 58 Wuichet, K. & Zhulin, I. B. Origins and diversification of a complex signal transduction system in prokaryotes. *Sci Signal* **3**, ra50, doi:10.1126/scisignal.2000724 (2010).
- 59 Briegel, A. *et al.* Universal architecture of bacterial chemoreceptor arrays. *Proc Natl Acad Sci U S A* **106**, 17181-17186 (2009).
- 60 Li, M. & Hazelbauer, G. L. Core unit of chemotaxis signaling complexes. *Proc Natl Acad Sci U S A* **108**, 9390-9395 (2011).
- 61 Briegel A, L. X., Bilwes AM, Hughes KT, Jensen GJ, Crane BR. Bacterial chemoreceptor arrays are hexagonally packed trimers of receptor dimers networked by rings of kinase and coupling proteins. *Proc Natl Acad Sci U S A* **109**, 3766-3771 (2012).
- 62 Liu, J. *et al.* Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of *Escherichia coli* minicells. *Proceedings of the National Academy of Sciences of the United States of America* **109**, E1481-E1488, doi:10.1073/pnas.1200781109 (2012).
- 63 Elliott, K. T., Zhulin, I. B., Stuckey, J. A. & DiRita, V. J. Conserved residues in the HAMP domain define a new family of proposed bipartite energy taxis receptors. *J Bacteriol* **191**, 375-387 (2009).
- 64 Jr., D. K. Bacterial chemotaxis in relation to neurobiology. *Ann Rev Neurosci* **3**, 43-75 (1980).
- 65 Ottemann KM, X. W., Shin YK, Koshland DE jr. A piston model for transmembrane signaling of the aspartate receptor. *Science* **285**, 1751-1754 (1999).
- 66 Parkinson, J. Signaling Mechanisms of HAMP Domains in Chemoreceptors and Sensor Kinases. *Annu Rev Microbiol* **64**, 101-122 (2010).
- 67 Gosink KK, Z. Y., Parkinson JS. Mutational Analysis of N381, a Key Trimer Contact Residue in Tsr, the *Escherichia coli* Serine Chemoreceptor. *J Bacteriol* **193** (2011).
- 68 Ames, P., Studdert, C. A., Relser, R. H. & Parkinson, J. S. Collaborative signaling by mixed chemoreceptor teams in *Escherichia coli*. *Proc Natl Acad Sci U S A* **99**, 7060-7065 (2002).
- 69 Studdert, C. A. & Parkinson, J. S. Crosslinking snapshots of bacterial chemoreceptor squads. *Proc Natl Acad Sci U S A* **101**, 2117-2122 (2004).

- 70 Natale AM, D. J., Piasta KN, Falke JJ. Structure, Function, and On-Off Switching of a Core Unit Contact between CheA Kinase and CheW Adaptor Protein in the Bacterial Chemosensory Array: A Disulfide Mapping and Mutagenesis Study. *Biochemistry* **52**, 7753-7765 (2013).
- 71 Swain KE, G. M., Falke JJ. Engineered Socket Study of Signaling through a 4-Helix Bundle: Evidence for a Yin-Yang Mechanism in the Kinase Control Module of the Aspartate Receptor. *Biochemistry* **48**, 9266-9277 (2009).
- 72 Li, M., Khursigara, C. M., Subramaniam, S. & Hazelbauer, G. L. Chemotaxis Kinase CheA is Activated by Three Neighboring Chemoreceptor Dimers as Effectively as by Receptor Clusters. *Mol Microbiol* **79**, 766-785 (2011).
- 73 Ortega, D. R. *et al.* Conformational Coupling between Receptor and Kinase Binding Sites through a Conserved Salt Bridge in a Signaling Complex Scaffold Protein. *PLoS Comput Biol* **9** (2013).
- 74 Vu A, W. X., Zhou H, Dahlquist FW. The Receptor-CheW Binding Interface in Bacterial Chemotaxis. *J Mol Biol* **415**, 759-767 (2012).
- 75 Wang X, V. A., Lee K, Dahlquist FW. CheA-Receptor Interaction Sites in Bacterial Chemotaxis. *J Mol Biol* **422**, 282-290 (2012).
- 76 Briegel A, O. D., Tocheva EI, Wuichet K, Li Z, Chen S, Muller A, Iancu CV, Murphy GE, Dobro MJ, Zhulin IB, Jensen GJ. Universal architecture of bacterial chemoreceptor arrays. *Proc Natl Acad Sci U S A* **106**, 17181-17186 (2009).
- 77 Airola MV, H. D., Sukomon N, Widom J, Sircar R, Borbat PP, Freed JH, Watts KJ, Crane BR. Architecture of the soluble receptor Aer2 indicates an in-line mechanism for PAS and HAMP domain signaling. *J Mol Biol* **425**, 886-901 (2013).
- 78 Li, X. *et al.* The 3.2 Angstrom Resolution Structure of a Receptor:CheA:CheW Signaling Complex Defines Overlapping Binding Sites and Key Residue Interactions within Bacterial Chemosensory Arrays. *Biochemistry* **52**, 3852-3865 (2013).
- 79 Liu J, H. B., Morado DR, Jani S, Manson MD, Margolin W. Molecular architecture of chemoreceptor arrays revealed by cryoelectron tomography of Escherichia coli minicells. *Proc Natl Acad Sci U S A* **109**, E1481-1488 (2012).
- 80 Ortega, D. R. *et al.* A phenylalanine rotameric switch for signal-state control in bacterial chemoreceptors. *Nat Commun* **4**, 2881 (2013).
- 81 Starrett DJ, F. J. Adaptation Mechanism of the Aspartate Receptor: Electrostatics of the Adaptation Subdomain Play a Key Role in Modulating Kinase Activity. *Biochemistry* **44**, 1550-1560 (2005).
- 82 Pedetta, A., Parkinson, J. S. & Studdert, C. A. Signalling-dependent interactions between the kinase-coupling protein CheW and chemoreceptors in living cells. *Mol Microbiol* **93**, 1144-1155 (2014).
- 83 Garvis S, M. A., Ball G, de Bentzmann S, Wiehlmann L, Ewbank JJ, Tummler B, Filloux A. Caenorhabditis elegans Semi-Automated Liquid Screen Reveals a Specialized Role for the Chemotaxis Gene cheB2 in Pseudomonas aeruginosa Virulence. *PLoS Pathogens* **5**, 13 (2009).
- 84 McLaughlin HP, C. D., McCarthy Y, Ryan RP, Dow JM. An orphan chemotaxis sensor regulates virulence and antibiotic tolerance in the human pathogen Pseudomonas aeruginosa. *PLoS ONE* **7**, 10 (2012).
- 85 Nishiyama S, S. D., Itoh Y, Suzuki K, Tajima H, Hyakutake A, Homma M, Butler-Wu SM, Camilli A, Kawagishi I. Mlp24 (McpX) of Vibrio cholerae implicated in pathogenicity functions as a chemoreceptor for multiple amino acids. *Infect Immun.* **80**, 9 (2012).
- 86 Rahman, H. *et al.* Characterisation of a Multi-ligand Binding Chemoreceptor CcmL (Tlp3) of Campylobacter jejuni. *PLoS Pathogens* **10** (2014).
- 87 Lertsethtakarn P, O. K., Hendrixson DR. Motility and chemotaxis in Campylobacter and Helicobacter. *Annu Rev Microbiol* **65**, 389-410 (2011).

- 88 Hagman KE, P. S., Popova TG, Norgard MV. Evidence for a methyl-accepting chemotaxis protein gene (mcp1) that encodes a putative sensory transducer in virulent *Treponema pallidum*. *Infect Immun* **65**, 1701 (1997).
- 89 Sze CW, Z. K., Kariu T, Pal U, Li C. *Borrelia burgdorferi* Needs Chemotaxis to Establish Infection in Mammals and To Accomplish Its Enzootic Cycle. *Infect Immun* **80** (2012).
- 90 Chaudhuri, R. R. & Henderson, I. R. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol* **12**, 214-226, doi:10.1016/j.meegid.2012.01.005 (2012).
- 91 Whitchurch CB, L. A., Young MD, Kennedy D, Sargent JL, Bertrand JJ, Semmier ABT, Mellick AS, Martin PR, Alm RA, Hobbs M, Beatson SA, Huang B, Nguyen L, Commolli JC, Engel JN, Darzins A, Mattick JS. Characterization of a complex chemosensory signal transduction system which controls twitching motility in *Pseudomonas aeruginosa*. *Mol Microbiol* **52**, 21 (2004).
- 92 Hickman JW, T. D., Harwood CS. A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. *Proc Natl Acad Sci U S A* **102**, 6 (2005).
- 93 Yang Z, G. Y., Xu D, Kaplan HB, Shi W. A new set of chemotaxis homologues is essential for *Myxococcus xanthus* social motility. *Mol Microbiol* **30**, 1123-1130 (1998).
- 94 Bustamante VH, M.-F. I., Vlamakis HC, Zusman DR. Analysis of the Frz signal transduction system of *Myxococcus xanthus* shows the importance of the conserved C-terminal region of the cytoplasmic chemoreceptor FrzCD in sensing signals. *Mol Microbiol* **53**, 1501-1513 (2004).
- 95 Burrows, L. L. *Pseudomonas aeruginosa* twitching motility: type IV pili in action. *Annu Rev Microbiol* **66**, 493-520 (2012).
- 96 Ryan, R. Cyclic di-GMP signalling and the regulation of bacterial virulence. *Microbiology* **159**, 1286-1297 (2013).
- 97 Gray, H. Anatomy of the Human Body. Publisher: Lea & Febiger (PA, USA). (1918).
- 98 Briegel, A. et al. Structure of bacterial cytoplasmic chemoreceptor arrays and implications for chemotactic signaling. *eLIFE* **3**, e02151 (2014).
- 99 Butler, S. M. & Camilli, A. Going against the grain: Chemotaxis and infection in *Vibrio cholerae*. *Nat Rev Microbiol* **3**, 611-620 (2005).
- 100 Zhu, S., Kojima, S. & Homma, M. Structure, gene regulation, and environmental response of flagella in *Vibrio*. *Front. Microbiol*, doi:10.3389/fmicb.2013.00410 (2013).
- 101 Martin, A. C., Wadhams, G. H. & Armitage, J. P. The roles of the multiple CheW and CheA homologues in chemotaxis and in chemoreceptor localization in *Rhodobacter sphaeroides*. *Mol Microbiol* **40**, 1261-1272 (2001).
- 102 Porter SL, W. G., Armitage JP. *Rhodobacter sphaeroides*: complexity in chemotactic signalling. *Trends in Microbiology* **16**, 251-260 (2008).
- 103 Wadhams, G. H., Martin, A. C. & Armitage, J. P. Identification and localization of a methyl-accepting chemotaxis protein in *Rhodobacter sphaeroides*. *Mol Microbiol* **36**, 1222-1233 (2000).
- 104 Roberts, M. A. J. et al. A model invalidation-based approach for elucidating biological signaling pathways, applied to the chemotaxis pathway in *R. sphaeroides*. *BMC Systems Biology* **3**, 105-118 (2009).
- 105 Hamer, R., Chen, P. Y., Armitage, J. P., Reinert, G. & Deane, C. M. Deciphering chemotaxis pathways using cross species comparisons. *BMC Systems Biology* **4**, 3-21 (2010).
- 106 Tran HT, K. J., Antommattei, Lovley DR, Weis RM. Comparative genomics of *Geobacter* chemotaxis genes reveals diverse signaling function. *BMC Genomics* **9**, 471-485 (2008).
- 107 Altschul, S. F. et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389-3402 (1997).

- 108 Alexander, R. P. & Zhulin, I. B. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **104**, 2885-2890 (2007).
- 109 Giamarellous, H. Therapeutic guidelines for *Pseudomonas aeruginosa* infections. *International Journal of Antimicrobial Agents* **16**, 4 (2000).
- 110 Ferrandez A, H. A., Summerfield DT, Harwood CS. Cluster II *che* genes from *Pseudomonas aeruginosa* are required for an optimal chemotactic response. *J Bacteriol* **184**, 10 (2002).
- 111 Guvener ZT, T. D., Harwood CS. Two different *Pseudomonas aeruginosa* chemosensory signal transduction complexes localize to cell poles and form and remould in stationary phase. *Mol Microbiol* **61**, 13 (2006).
- 112 Kato J, N. T., Kuroda A, Ohtake H. Cloning and characterization of chemotaxis genes in *Pseudomonas aeruginosa*. *Biosci. Biotechnol. Biochem.* **63**, 7 (1999).
- 113 Kim HE, S. M., Kuroda A, Takiguchi N, Ohtake H, Kato J. Identification and characterization of the chemotactic transducer in *Pseudomonas aeruginosa* PAO1 for positive chemotaxis to trichloroethylene. *J Bacteriol* **188**, 3 (2006).
- 114 Kuroda A, K. T., Taguchi K, Nikata T, Kato J, Ohtake H. Molecular cloning and characterization of a chemotactic transducer gene in *Pseudomonas aeruginosa*. *J Bacteriol* **177**, 7 (1995).
- 115 Taguchi K, F. H., Kuroda A, Kato J, Ohtake H. Genetic identification of chemotactic transducers for amino acids in *Pseudomonas aeruginosa*. *1997* **143**, 7 (1997).
- 116 O'Connor JR, K. N., Huangyutitham V, Wiggins PA, Harwood CS. Surface sensing and lateral subcellular localization of WspA, the receptor in a chemosensory-like system leading to c-di-GMP production. *Mol Microbiol* **86**, 10 (2012).
- 117 Hong CS, S. M., Kuroda A, Ikeda T, Takiguchi N, Ohtake H, Kato J. Chemotaxis proteins and transducers for aerotaxis in *Pseudomonas aeruginosa*. *FEMS Microbiology Letters* **231**, 6 (2004).
- 118 Alvarez-Ortega C, H. C. Identification of a malate chemoreceptor in *Pseudomonas aeruginosa* by screening for chemotaxis defects in an energy taxis-deficient mutant. *Applied and Environmental Microbiology* **73**, 3 (2007).
- 119 Liu X, W. P., Parales JV, Parales RE. Chemotaxis to pyrimidines and identification of a cytosine chemoreceptor in *Pseudomonas aeruginosa*. *J Bacteriol* **191**, 8 (2009).
- 120 Morgan R, K. S., Hwang SH, Hassett DJ, Sauer K. BdlA, a chemotaxis regulator essential for biofilm dispersion in *Pseudomonas aeruginosa*. *J Bacteriol* **188**, 9 (2006).
- 121 Bardy SL, M. J. Polar localization of a soluble methyl-accepting protein of *Pseudomonas aeruginosa*. *J Bacteriol* **187**, 5 (2005).
- 122 Mauriello, E. M. F., Astling, D. P., Sliusarenko, O. & Zusman, D. R. Localization of a bacterial cytoplasmic receptor is dynamic and changes with cell-cell contacts. *Proc Natl Acad Sci U S A* **106**, 4852-4857 (2009).
- 123 Moine A, A. R., Espinosa L, Kirby JR, Zusman DR, Mignot T, Mauriello EMF. Functional organization of a multimodular bacterial chemosensory apparatus. *PLoS Genetics* **10** (2014).
- 124 Meier, V. M., Muschler, P. & Scharf, B. E. Functional analysis of nine putative chemoreceptor proteins in *Sinorhizobium meliloti*. *J Bacteriol* **189**, 1816-1827 (2007).
- 125 Meier, V. M. & Scharf, B. E. Cellular localization of predicted transmembrane and soluble chemoreceptors in *Sinorhizobium meliloti*. *191* **18** (2009).
- 126 Yoshihara S, S. F., Fujita H, Geng XX, Ikeuchi M. Novel putative photoreceptor and regulatory genes required for the positive phototactic movement of the unicellular motile cyanobacterium *Synechocystis* sp. PCC 6803. *Plant Cell Physiol* **41**, 1299-1304 (2000).

- 127 Bhaya D, T. A., Grossman AR. Light regulation of type IV pilus-dependent motility by chemosensor-like elements in *Synechocystis* PCC6803. *Proc Natl Acad Sci U S A* **98**, 7540-7545 (2001).
- 128 Wuichet, K. & Zhulin, I. B. Molecular evolution of sensory domains in cyanobacterial chemoreceptors. *Trends in Microbiology* **11**, 200-203 (2003).
- 129 Adler, J. CHEMOTAXIS IN BACTERIA. *Annu. Rev. Biochem.* **44**, 341-356 (1975).
- 130 Sourjik V, W. N. Responding to Chemical Gradients: Bacterial Chemotaxis. *Curr Opin Cell Biol.* **24**, 262-268 (2012).
- 131 Hazelbauer, G. L., Falke, J. J. & Parkinson, J. S. Bacterial chemoreceptors: high-performance signaling in networked arrays. *Trends Biochem.Sci.* **33**, 9-19 (2008).
- 132 Howitt, M. R. *et al.* ChePep Controls *Helicobacter pylori* Infection of the Gastric Glands and Chemotaxis in the Epsilonproteobacteria. *Mbio* **2**, doi:e00098-11 (2011).
- 133 Rader, B. A. *et al.* *Helicobacter pylori* perceives the quorum-sensing molecule AI-2 as a chemorepellent via the chemoreceptor TlpB. *Microbiology-Sgm* **157**, 2445-2455, doi:10.1099/mic.0.049353-0 (2011).
- 134 Rolig, A. S., Carter, J. E. & Ottemann, K. M. Bacterial chemotaxis modulates host cell apoptosis to establish a T-helper cell, type 17 (Th17)-dominant immune response in *Helicobacter pylori* infection. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 19749-19754, doi:10.1073/pnas.104598108 (2011).
- 135 Schweinitzer, T. & Josenhans, C. Bacterial energy taxis: a global strategy? *Archives of Microbiology* **192**, 507-520, doi:10.1007/s00203-010-0575-7 (2010).
- 136 Antunez-Lamas, M. *et al.* Role of motility and chemotaxis in the pathogenesis of *Dickeya dadantii* 3937 (ex *Erwinia chrysanthemi* 3937). *Microbiology-Sgm* **155**, 434-442, doi:10.1099/mic.0.022244-0 (2009).
- 137 Spagnuolo, A. M., DiRita, V. & Kirschner, D. A model for *Vibrio cholerae* colonization of the human intestine. *Journal of Theoretical Biology* **289**, 247-258, doi:10.1016/j.jtbi.2011.08.028 (2011).
- 138 Li, C. H. *et al.* Asymmetrical flagellar rotation in *Borrelia burgdorferi* nonchemotactic mutants. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 6169-6174, doi:10.1073/pnas.092010499 (2002).
- 139 Lux, R., Miller, J. N., Park, N. H. & Shi, W. Y. Motility and chemotaxis in tissue penetration of oral epithelial cell layers by *Treponema denticola*. *Infection and Immunity* **69**, 6276-6283, doi:10.1128/iai.69.10.6276-6283.2001 (2001).
- 140 Motaleb, M. A., Miller, M. R., Bakker, R. G., Li, C. H. & Charon, N. W. in *Two-Component Signaling Systems, Pt A Vol. 422 Methods in Enzymology* (eds M. I. Simon, B. R. Crane, & A. Crane) 421-+ (2007).
- 141 Maddock JR, S. L. Polar Location of the Chemoreceptor Complex in the *Escherichia coli* Cell. *Science* **259**, 1717-1723 (1993).
- 142 Greenfield, D. *et al.* Self-Organization of the *Escherichia coli* Chemotaxis Network Imaged with Super-Resolution Light Microscopy. *Plos Biology* **7**, doi:e1000137 (2009).
- 143 Kim, C., Jackson, M., Lux, R. & Khan, S. Determinants of chemotactic signal amplification in *Escherichia coli*. *Journal Of Molecular Biology* **307**, 119-135 (2001).
- 144 Kim, S. H., Wang, W. R. & Kim, K. K. Dynamic and clustering model of bacterial chemotaxis receptors: Structural basis for signaling and high sensitivity. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **99**, 11611-11615 (2002).
- 145 Goldman, J. P., Levin, M. D. & Bray, D. Signal amplification in a lattice of coupled protein kinases. *Molecular Biosystems* **5**, 1853-1859, doi:10.1039/b903397a (2009).
- 146 Briegel, A. *et al.* Location and architecture of the *Caulobacter crescentus* chemoreceptor array. *Molecular Microbiology* **69**, 30-41 (2008).

- 147 Khursigara, C. M., Wu, X. W. & Subramaniam, S. Chemoreceptors in *Caulobacter crescentus*: Trimers of receptor dimers in a partially ordered hexagonally packed array. *Journal Of Bacteriology* **190**, 6805-6810 (2008).
- 148 Zhang, P. J., Khursigara, C. M., Hartnell, L. M. & Subramaniam, S. Direct visualization of *Escherichia coli* chemotaxis receptor arrays using cryo-electron microscopy. *Proceedings Of The National Academy Of Sciences Of The United States Of America* **104**, 3777-3781 (2007).
- 149 Khursigara, C. M. *et al.* Lateral density of receptor arrays in the membrane plane influences sensitivity of the *E. coli* chemotaxis response. *Embo Journal* **30**, 1719-1729, doi:10.1038/emboj.2011.77 (2011).
- 150 Briegel, A., Beeby, M., Thanbichler, M. & Jensen, G. J. Activated chemoreceptor arrays remain intact and hexagonally packed. *Molecular Microbiology* **82**, 748-757, doi:10.1111/j.1365-2958.2011.07854.x (2011).
- 151 Falke, J. J. & Hazelbauer, G. L. Transmembrane signaling in bacterial chemoreceptors. *Trends Biochem. Sci.* **26**, 257-265 (2001).
- 152 Zhulin, I. B. The superfamily of chemotaxis transducers: from physiology to genomics and back. *Adv Microb Physiol* **45**, 157-198 (2001).
- 153 Alexander RP, Z. I. Evolutionary genomics reveals conserved structural determinants of signaling and adaptation in microbial chemoreceptors. *Proc Natl Acad Sci U S A* **104**, 2885-2890 (2007).
- 154 Hulko, M. *et al.* The HAMP domain structure implies helix rotation in transmembrane signaling. *Cell* **126**, 929-940 (2006).
- 155 Buron-Barral, M. D., Gosink, K. K. & Parkinson, J. S. Loss- and gain-of-function mutations in the F1-HAMP region of the *Escherichia coli* aerotaxis transducer Aer. *Journal Of Bacteriology* **188**, 3477-3486 (2006).
- 156 Mehan, R. S., White, N. C. & Falke, J. J. Mapping out regions on the surface of the aspartate receptor that are essential for kinase activation. *Biochemistry* **42**, 2952-2959 (2003).
- 157 Park, S. Y., Borbat, P.P., Gonzalez-Bonet, G., Bhatnagar, J., Freed, J.H., Bilwes, A.M., Crane, B.R. Reconstruction of the chemotaxis receptor:kinase assembly. *Nat. Struct. Mol. Biol.* **13**, 400-407 (2006).
- 158 Pollard, A. M., Bilwes, A. M. & Crane, B. R. The Structure of a Soluble Chemoreceptor Suggests a Mechanism for Propagating Conformational Signals. *Biochemistry* **48**, 1936-1944, doi:10.1021/bi801727m (2009).
- 159 Borkovich, K. A., Alex, L. A. & Simon, M. I. Attenuation of sensory receptor signaling by covalent modification. *Proc. Natl. Acad. Sci. U S A* **89**, 6756-6760 (1992).
- 160 Li, G., and Weis, R. M. Covalent modification regulates ligand binding to receptor complexes in the chemosensory system of *Escherichia coli*. *Cell* **100**, 357-365 (2000).
- 161 Bornhorst, J. A. & Falke, J. J. Attractant regulation of the aspartate receptor-kinase complex: Limited cooperative interactions between receptors and effects of the receptor modification state. *Biochemistry* **39**, 9486-9493 (2000).
- 162 Chao, X. *et al.* A receptor-modifying deamidase in complex with a signaling phosphatase reveals reciprocal regulation. *Cell* **124**, 561-571 (2006).
- 163 Hess, J. F., Bourret, R.B., & Simon, M.I. Histidine phosphorylation and phosphoryl group transfer in bacterial chemotaxis. *Nature* **336**, 139-143 (1988).
- 164 Kofoed, E. C., and Parkinson, J. S. Transmitter and receiver modules in bacterial signaling proteins. *Proc Natl Acad Sci U S A* **85**, 4981-4985 (1988).
- 165 Alex, L. A. & Simon, M. I. Protein histidine kinases and signal transduction in prokaryotes and eukaryotes. *Trends Genet.* **10**, 133-138 (1994).

- 166 Welch, M., Chinardet, N., Mourey, L., Birck, C. & Samama, J. P. Structure of the CheY-binding domain of histidine kinase CheA in complex with CheY. *Nature Struct. Biol.* **5**, 25-29 (1998).
- 167 Mourey, L. *et al.* Crystal structure of the CheA histidine phosphotransfer domain that mediates response regulator phosphorylation in bacterial chemotaxis. *J. Biol. Chem.* **276**, 31074-31082 (2001).
- 168 Quezada, C. M., Gradinaru, C., Simon, M. I., Bilwes, A. M. & Crane, B. R. Helical shifts generate two distinct conformers in the atomic resolution structure of the CheA phosphotransferase domain from *Thermotoga maritima*. *Journal of Molecular Biology* **341**, 1283-1294 (2004).
- 169 McEvoy, M. M., Muhandiram, D. R., Kay, L. E. & Dahlquist, F. W. Structure and dynamics of a CheY-binding domain of the chemotaxis kinase CheA determined by nuclear magnetic resonance spectroscopy. *Biochemistry* **35**, 5633-5640 (1996).
- 170 Park, S. Y., Beel, B. D., Simon, M. I., Bilwes, A. M. & Crane, B. R. In different organisms, the mode of interaction between two signaling proteins is not necessarily conserved. *Proc Natl Acad Sci U S A* **101**, 11646-11651 (2004).
- 171 Wuichet, K., and Alexander, R.P., and Zhulin, I.B. Comparative genomic and protein sequence analyses of a complex system controlling bacterial chemotaxis. *Meth. Enzymol.* **422**, 3-31 (2007).
- 172 Sanders, D. A., Mendez, B. & Koshland, D. E. J. Role of the CheW protein in bacterial chemotaxis: overexpression is equivalent to absence. *J. Bacteriol.* **171**, 6271-6278 (1989).
- 173 Gegner, J. A. & Dahlquist, F. W. Signal transduction in bacteria: CheW forms a reversible complex with the protein kinase CheA. *Proc. Natl. Acad. Sci. U S A* **88**, 750-754 (1991).
- 174 Griswold, I. J. *et al.* The solution structure and interactions of CheW from *Thermotoga maritima*. *Nature Struct. Biol.* **9**, 567-568 (2002).
- 175 Hamel, D. J. & Dahlquist, F. W. The contact interface of a 120 kD CheA-CheW complex by methyl TROSY interaction spectroscopy. *J Am Chem Soc* **127**, 9676-9677 (2005).
- 176 Miller, A. S., Kohout, S. C., Gilman, K. A. & Falke, J. J. CheA kinase of bacterial chemotaxis: Chemical mapping of four essential docking sites. *Biochemistry* **45**, 8699-8711 (2006).
- 177 Zhao, J. H. & Parkinson, J. S. Mutational analysis of the chemoreceptor-coupling domain of the Escherichia coli chemotaxis signaling kinase CheA. *Journal Of Bacteriology* **188**, 3299-3307 (2006).
- 178 Zhao, J. S. & Parkinson, J. S. Cysteine-scanning analysis of the chemoreceptor-coupling domain of the Escherichia coli chemotaxis signaling kinase CheA. *Journal Of Bacteriology* **188**, 4321-4330 (2006).
- 179 Asinas, A. E. & Weis, R. M. Competitive and cooperative interactions in receptor signaling complexes. *Journal Of Biological Chemistry* **281**, 30512-30523 (2006).
- 180 Cardozo, M. J., Massazza, D. A., Parkinson, J. S. & Studdert, C. A. Disruption of chemoreceptor signalling arrays by high levels of CheW, the receptor-kinase coupling protein. *Molecular Microbiology* **75**, 1171-1181, doi:10.1111/j.1365-2958.2009.07032.x (2010).
- 181 Ames, P. & Parkinson, J. S. Constitutively signaling fragments of Tsr, the *E.coli* serine chemoreceptor. *J. Bacteriol.* **176**, 6340-6348 (1994).
- 182 Borkovich, K. A., Kaplan, N., Hess, J. F. & Simon, M. I. Transmembrane signal transduction in bacterial chemotaxis involves ligand-dependent activation of phosphate group transfer. *Proc. Natl. Acad. Sci. U S A* **86**, 1208-1212 (1989).
- 183 Wang, X. Q., Wu, C., Anh Vu, J. E. S. & Dahlquist, F. W. Computational and Experimental Analyses Reveal the Essential Roles of Interdomain Linkers in the

- Biological Function of Chemotaxis Histidine Kinase CheA. *J. Am. Chem. Soc.* **134**, 16107-16110, doi:10.1021/ja3056694 (2012).
- 184 Otwinowski, A. & Minor, W. Processing of X-ray diffraction data in oscillation mode. *Methods Enzymol.* **276**, 307-325 (1997).
- 185 McCoy, A. J. *et al.* Phaser crystallographic software. *Journal Of Applied Crystallography* **40**, 658-674 (2007).
- 186 McRee, D. E. XtalView Xfit - A versatile program for manipulating atomic coordinates and electron density. *Journal of Structural Biology* **125**, 156-165, doi:10.1006/jsbi.1999.4094 (1999).
- 187 Adams, P. D. *et al.* The Phenix software for automated determination of macromolecular structures. *Methods* **55**, 94-106, doi:10.1016/j.ymeth.2011.07.005 (2011).
- 188 Su, J. *et al.* Crystal structure of a novel non-Pfam protein PF2046 solved using low resolution B-factor sharpening and multi-crystal averaging methods. *Protein Cell* **1**, 453-458, doi:10.1007/s13238-010-0045-7 (2010).
- 189 Altschul SF, M. T., Schaffer AA, Zhang Z, Miller W, Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* **25**, 14 (1997).
- 190 Katoh, K. & Toh, H. Parallelization of the MAFFT multiple sequence alignment program. *Bioinformatics* **26**, 1899-1900, doi:10.1093/bioinformatics/btq224 (2010).
- 191 Tamura K, P. D., Peterson N, Stecher G, Nei M, Kumar S. MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. *Molecular Biology and Evolution* **28**, 9 (2011).
- 192 Moreno-Hagelsieb, G. & Collado-Vides, J. A powerful non-homology method for the prediction of operons in prokaryotes. *Bioinformatics* **18 Suppl 1**, S329-336 (2002).
- 193 da Silveira, C. H. *et al.* Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. *Proteins* **74**, 727-743, doi:10.1002/prot.22187 (2009).
- 194 Wang, X. Q., Vu, A., Lee, K. & Dahlquist, F. W. CheA-Receptor Interaction Sites in Bacterial Chemotaxis. *Journal of Molecular Biology* **422**, 282-290, doi:10.1016/j.jmb.2012.05.023 (2012).
- 195 Park, S. Y., Quezada, C. M., Bilwes, A. M. & Crane, B. R. Subunit exchange by CheA histidine kinases from the mesophile *Escherichia coli* and the thermophile *thermotoga maritima*. *Biochemistry* **43**, 2228-2240 (2004).
- 196 Berggard, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**, 2833-2842, doi:10.1002/pmic.200700131 (2007).
- 197 Mowery, P., Ostler, J. B. & Parkinson, J. S. Different Signaling Roles of Two Conserved Residues in the Cytoplasmic Hairpin Tip of Tsr, the *Escherichia coli* Serine Chemoreceptor. *Journal of Bacteriology* **190**, 8065-8074, doi:10.1128/jb.01121-08 (2008).
- 198 Boukhvalova, M., Dahlquist, F. W. & Stewart, R. C. CheW binding interactions with CheA and Tar - Importance for chemotaxis signaling in *Escherichia coli*. *J. Biol. Chem.* **277**, 22251-22259 (2002).
- 199 Boukhvalova, M., VanBruggen, R. & Stewart, R. C. CheA kinase and chemoreceptor interactions on CheW. *J. Biol. Chem.* **277**, 23596-23603 (2002).
- 200 Liu, J. D. & Parkinson, J. S. Genetic evidence for interaction between the CheW and Tsr proteins during chemoreceptor signaling by *Escherichia coli*. *J. Bacteriol* **173**, 4941-4951 (1991).
- 201 Weigt, M., White, R. A., Szurmant, H., Hoch, J. A. & Hwa, T. Identification of direct residue contacts in protein-protein interaction by message passing. *Proc Natl Acad Sci U S A* **106**, 67-72, doi:10.1073/pnas.0805923106 (2009).

- 202 Lybarger, S. R., Nair, U., Lilly, A. A., Hazelbauer, G. L. & Maddock, J. R. Clustering
requires modified methyl-accepting sites in low-abundance but not high-abundance
chemoreceptors of *Escherichia coli*. *Molecular Microbiology* **56**, 1078-1086 (2005).
- 203 Watts, K. J., Taylor, B. L. & Johnson, M. S. PAS/poly-HAMP signalling in Aer-2, a
soluble haem-based sensor. *Molecular Microbiology* **79**, 686-699, doi:10.1111/j.1365-
2958.2010.07477.x (2011).
- 204 Gibbons, D. L. *et al.* Conformational change and protein protein interactions of the fusion
protein of Semliki Forest virus. *Nature* **427**, 320-325, doi:10.1038/nature02239 (2004).
- 205 Igonet, S. *et al.* X-ray structure of the arenavirus glycoprotein GP2 in its postfusion
hairpin conformation. *Proceedings of the National Academy of Sciences of the United
States of America* **108**, 19967-19972, doi:10.1073/pnas.1108910108 (2011).
- 206 Lamb, R. A. & Jardetzky, T. S. Structural basis of viral invasion: lessons from
paramyxovirus F. *Curr. Opin. Struct. Biol.* **17**, 427-436, doi:10.1016/j.sbi.2007.08.016
(2007).
- 207 Luque, L. E. & Russell, C. J. Spring-loaded heptad repeat residues regulate the
expression and activation of paramyxovirus fusion protein. *J. Virol.* **81**, 3130-3141,
doi:10.1128/jvi.02464-06 (2007).
- 208 Underbakke, E. S., Zhu, Y. M. & Kiessling, L. L. Protein footprinting in a complex
milieu: identifying the interaction surfaces of the chemotaxis adaptor protein CheW. *J.
Mol. Biol.* **409**, 483-495 (2011).
- 209 Levit, M. N., Grebe, T.W., Stock, J.B. Organization of the receptor-kinase signaling array
that regulates *Escherichia coli* chemotaxis. *J. Biol. Chem.* **277**, 36748-36754 (2002).
- 210 Hartley-Tassell, L. E. *et al.* Identification and characterization of the aspartate
chemosensory receptor of *Campylobacter jejuni*. *Mol Microbiol* **75**, 710-730 (2010).
- 211 Richardson JF, F. J., Kramer JM, Thwaites RT, Bolton FJ, Wareing DRA, Gordon JA.
Coinfection with *Campylobacter* species: an epidemiological problem? *Journal of
Applied Microbiology* **91**, 206-211 (2001).
- 212 Soding J, B. A., Lupas AN. The HHpred interactive server for protein homology detection
and structure prediction. *Nucleic Acids Research* **33**, 5 (2005).
- 213 Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M. & Barton, G. J. Jalview
Version 2-a multiple sequence alignment editor and analysis workbench. *Bioinformatics*
25, 1189-1191, doi:DOI 10.1093/bioinformatics/btp033 (2009).
- 214 Patskovsky Y, O. S., Freeman J, Hu S, Smith D, Bain K, Wasserman SR, Sauder JM,
Burley SK, Almo SC. Crystal structure of Mcp_N and cache domains of methyl-accepting
chemotaxis protein from *Vibrio cholerae*. doi:10.2210/pdb3c8c/pdb (2008).
- 215 Ouali M, K. R. Cascaded multiple classifiers for secondary structure prediction. *Protein
Science* **9**, 15 (2000).
- 216 Rost, B. Protein secondary structure prediction continues to rise. *Journal of Structural
Biology* **134**, 15 (2001).
- 217 Jones, D. Protein secondary structure prediction based on position-specific scoring
matrices. *Journal of Molecular Biology* **292**, 8 (1999).
- 218 Glekas GD, M. B., Kroc A, Duelfer KA, Lei V, Rao CV, Ordal GW. The *Bacillus subtilis*
chemoreceptor McpC senses multiple ligands using two discrete mechanisms. *J Biol
Chem* **287**, 7 (2012).
- 219 Muller J, S. S., Ordal GW, Saxild HH. Functional and genetic characterization of mcpC,
which encodes a third methyl-accepting chemotaxis protein in *Bacillus subtilis*.
Microbiology **143**, 10 (1997).
- 220 Taguchi K, F. H., Kuroda A, Kato J, Ohtake H. Genetic identification of chemotactic
transducers for amino acids in *Pseudomonas aeruginosa*. *Microbiology* **143**, 7 (1997).

- 221 Rico-Jimenez, M. *et al.* Paralogous chemoreceptors mediate chemotaxis towards protein amino acids and the non-protein amino acid gamma-aminobutyrate (GABA). *Mol Microbiol* **88**, 14 (2013).
- 222 Oku S, K. A., Tajima T, Nakashimada Y, Kato J. Identification of Chemotaxis Sensory Proteins for Amino Acids in *Pseudomonas fluorescens* Pf0-1 and Their Involvement in Chemotaxis to Tomato Root Exudate and Root Colonization. *Microbes Environ.* **27**, 8 (2012).
- 223 Webb BA, H. S., Helm RF, Scharf BE. Sinorhizobium meliloti Chemoreceptor McpU Mediates Chemotaxis toward Host Plant Exudates through Direct Proline Sensing. *Appl Environ Microbiol* **80** (2014).
- 224 Porter, S. L., Wadhams, G. H. & Armitage, J. P. Signal processing in complex chemotaxis pathways. *Nat Rev Microbiol* **9**, 153-165 (2011).
- 225 Bischoff DS, O. G. *Bacillus subtilis* chemotaxis: a deviation from the Escherichia coli paradigm. *Mol Microbiol* **6**, 23-28 (1992).
- 226 Adler, J. Chemotaxis in bacteria. *J Supramol Struct* **4**, 305-317 (1976).
- 227 Parkinson, J. Genetics of chemotactic behavior in bacteria. *Cell* **4**, 183-188 (1975).
- 228 Lybarger SR, M. J. Differences in the polar clustering of the high- and low-abundance chemoreceptors of Escherichia coli. *Proc Natl Acad Sci U S A* **97**, 8057-8062 (2000).
- 229 Stover CK, P. X., Erwin AL, Mizoguchi SD, Warren P, Hickey MJ, Brinkman FSL, Hufnagle WO, Kowalik DJ, Lagrou M, Garber RL, Goltry L, Tolentino E, Westbrook-Wadman S, Yuan Y, Brody LL, Coulter SN, Folger KR, Kas A, Larbig K, Lim R, Smith K, Spencer D, Wong GKS, Wu Z, Paulsen IT, Reizer J, Saier MH, Hancock REW, Lory S, Olson MV. Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**, 6 (2000).
- 230 Masduki A, N. J., Ohga T, Umezaki R, Kato J, Ohtake K. Isolation and characterization of chemotaxis mutants and genes of *Pseudomonas aeruginosa*. *J Bacteriol* **177**, 5 (1995).
- 231 Barraud N, S. D., Klebensberger J, Webb JS, Hassett DJ, Rice SA, Kjelleberg S. Nitric oxide signaling in *Pseudomonas aeruginosa* biofilms mediates phosphodiesterase activity, decreased cyclic di-GMP levels, and enhanced dispersal. *J Bacteriol* **191**, 10 (2009).
- 232 Schuster M, H. A., Harwood CS, Greenberg EP. The *Pseudomonas aeruginosa* RpoS regulon and its relationship to quorum sensing. *Mol Microbiol* **51**, 973-985 (2004).
- 233 Kato J, K. H., Takiguchi N, Kuroda A, Ohtake H. *Pseudomonas aeruginosa* as a Model Microorganism for Investigation of Chemotactic Behaviors in Ecosystem. *Journal of Bioscience and Bioengineering* **106**, 1-7 (2008).
- 234 Sampedro I, P. R., Krell T, Hill JE. *Pseudomonas* Chemotaxis. *FEMS Microbiol Rev.*, doi:10.1111/1574-6976.10281 (2014).
- 235 Kim HE, S. M., Kuroda A, Takiguchi N, Ohtake H, Kato J. Ethylene Chemotaxis in *Pseudomonas aeruginosa* and Other *Pseudomonad* Species. *Microbes Environ.* **22**, 186-189 (2007).
- 236 Vangnai AS, T. K., Oku S, Kataoka N, Nitisakulkan T, Tajima T, Kato J. Identification of CtpL as a chromosomally encoded chemoreceptor for 4-chloroaniline and catechol in *Pseudomonas aeruginosa* PAO1. *Appl Environ Microbiol* **79**, 7241-7248 (2013).
- 237 Rodionov DA, D. I., Arkin AP, Alm EJ, Gelfand MS. Dissimilatory Metabolism of Nitrogen Oxides in Bacteria: Comparative Reconstruction of Transcriptional Networks. *PLoS Comput Biol* **1** (2005).
- 238 Garnacho-Montero, J. & Amaya-Villar, R. Multiresistant *Acinetobacter baumannii* infections: epidemiology and management. *Curr Opin Infect Dis* **23**, 332-339 (2010).
- 239 Bitrian, M., Gonzalez, R. H., Paris, G., Hellingwerf, K. J. & Nudel, C. B. Blue-light-dependent inhibition of twitching motility in *Acinetobacter baylyi* ADP1: additive

- involvement of three BLUF-domain-containing proteins. *Microbiology* **159**, 1828-1841 (2013).
- 240 Salunkhe P, S. C., Morgan JAW, Panagea S, Walshaw MJ, Hart CA, Geffers R, Tumbler B, Winstanley C. A cystic fibrosis epidemic strain of *Pseudomonas aeruginosa* displays enhanced virulence and antimicrobial resistance. *J Bacteriol* **187**, 13 (2005).
- 241 Roy PH, T. S., Larouche A, Elbourne L, Tremblay S, Ren Q, Dodson R, Harkins D, Shay R, Watkins K, Mahamoud Y, Paulsen IT. Complete genome sequence of the multiresistant taxonomic outlier *Pseudomonas aeruginosa* PA7. *PLoS ONE* **5**, 10 (2010).
- 242 Borziak, K., Fleetwood, A. D. & Zhulin, I. B. Chemoreceptor Gene Loss and Acquisition via Horizontal Gene Transfer in *Escherichia coli*. *J Bacteriol* **195** (2013).
- 243 Lane MC, L. A., Markyvech TA, Hagan EC, Mobley HLT. Uropathogenic *Escherichia coli* Strains Generally Lack Functional Trg and Tap Chemoreceptors Found in the Majority of *E. coli* Strains Strictly Residing in the Gut. *J Bacteriol* **188** (2006).
- 244 Kaper, J. B., Nataro, J. P. & Mobley, H. L. T. Pathogenic *Escherichia coli*. *Nat Rev Microbiol* **2**, 123-140, doi:10.1038/nrmicro818 (2004).
- 245 Croxen, M. A. & Finlay, B. B. Molecular mechanisms of *Escherichia coli* pathogenicity. *Nat Rev Microbiol* **8**, 26-38, doi:10.1038/nrmicro2265 (2010).
- 246 Giron, J. A., Torres, A. G., Freer, E. & Kaper, J. B. The flagella of enteropathogenic *Escherichia coli* mediate adherence to epithelial cells. *Mol Microbiol* **44**, 361-379 (2002).
- 247 Lane, M. C. *et al.* Role of motility in the colonization of uropathogenic *Escherichia coli* in the urinary tract. *Infection and Immunity* **73**, 7644-7656, doi:10.1128/IAI.73.11.7644-7656.2005 (2005).
- 248 Hegde, M. *et al.* Chemotaxis to the quorum-sensing signal AI-2 requires the Tsr chemoreceptor and the periplasmic LsrB AI-2-binding protein. *J Bacteriol* **193**, 768-773, doi:10.1128/JB.01196-10 (2011).
- 249 Springer, M. S., Goy, M. F. & Adler, J. Sensory transduction in *Escherichia coli*: two complementary pathways of information processing that involve methylated proteins. *Proc Natl Acad Sci U S A* **74**, 3312-3316 (1977).
- 250 Greer-Phillips, S. E., Alexandre, G., Taylor, B. L. & Zhulin, I. B. Aer and Tsr guide *Escherichia coli* in spatial gradients of oxidizable substrates. *Microbiology* **149**, 2661-2667 (2003).
- 251 Hazelbauer, G. L. Maltose chemoreceptor of *Escherichia coli*. *J Bacteriol* **122**, 206-214 (1975).
- 252 Tso, W. W. & Adler, J. Negative chemotaxis in *Escherichia coli*. *J Bacteriol* **118**, 560-576 (1974).
- 253 Harayama, S., Palva, E. T. & Hazelbauer, G. L. Transposon-insertion mutants of *Escherichia coli* K12 defective in a component common to galactose and ribose chemotaxis. *Mol Gen Genet* **171**, 193-203 (1979).
- 254 Frye, J. *et al.* Identification of new flagellar genes of *Salmonella enterica* serovar Typhimurium. *J Bacteriol* **188**, 2233-2243, doi:10.1128/JB.188.6.2233-2243.2006 (2006).
- 255 Jaureguy, F. *et al.* Phylogenetic and genomic diversity of human bacteremic *Escherichia coli* strains. *BMC Genomics* **9**, 560, doi:10.1186/1471-2164-9-560 (2008).
- 256 Touchon, M. *et al.* Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. *Plos Genetics* **5**, e1000344, doi:10.1371/journal.pgen.1000344 (2009).
- 257 Wirth, T. *et al.* Sex and virulence in *Escherichia coli*: an evolutionary perspective. *Mol Microbiol* **60**, 1136-1151, doi:10.1111/j.1365-2958.2006.05172.x (2006).
- 258 Ochman, H. & Selander, R. K. Standard Reference Strains of *Escherichia-Coli* from Natural-Populations. *J Bacteriol* **157**, 690-693 (1984).

- 259 Escobar-Paramo, P. *et al.* A specific genetic background is required for acquisition and
expression of virulence factors in *Escherichia coli*. *Molecular Biology and Evolution* **21**,
1085-1094, doi:DOI 10.1093/molbev/msh118 (2004).
- 260 Chaudhuri RR, H. I. The evolution of the *Escherichia coli* phylogeny. *Infect Genet Evol*
12, 214-226 (2012).
- 261 Zhang, Y. & Lin, K. A phylogenomic analysis of *Escherichia coli* / *Shigella* group:
implications of genomic features associated with pathogenicity and ecological
adaptation. *BMC evolutionary biology* **12**, 174, doi:10.1186/1471-2148-12-174 (2012).
- 262 Skippington, E. & Ragan, M. A. Phylogeny rather than ecology or lifestyle biases the
construction of *Escherichia coli*-*Shigella* genetic exchange communities. *Open biology* **2**,
120112, doi:10.1098/rsob.120112 (2012).
- 263 Katoh, K. & Toh, H. Recent developments in the MAFFT multiple sequence alignment
program. *Brief Bioinform* **9**, 286-298, doi:Doi 10.1093/Bib/Bbn013 (2008).
- 264 Tamura, K., Dudley, J., Nei, M. & Kumar, S. MEGA4: Molecular evolutionary genetics
analysis (MEGA) software version 4.0. *Molecular Biology and Evolution* **24**, 1596-1599,
doi:DOI 10.1093/molbev/msm092 (2007).
- 265 Guindon, S., Dufayard, J. F., Hordijk, W., Lefort, V. & Gascuel, O. PhyML: Fast and
Accurate Phylogeny Reconstruction by Maximum Likelihood. *Infect Genet Evol* **9**, 384-
385 (2009).
- 266 Camacho, C. *et al.* BLAST plus : architecture and applications. *BMC Bioinformatics* **10**,
doi:Artn 421
Doi 10.1186/1471-2105-10-421 (2009).
- 267 Wattam, A. R., Abraham, D., Dalay, O., Disz, T. L. & Sobral, B. W. PATRIC, the bacterial
bioinformatics database and analysis resource. *Nucleic Acids Research* **42**, D581-391
(2014).
- 268 Reddy, T. B. K. *et al.* The Genomes OnLine Database (GOLD) v.5: a metadata
management system based on a four level (meta)genome project classification. *Nucleic
Acids Research*, doi:10.1093/nar/gku950 (2014).
- 269 Miquel, S. *et al.* Complete genome sequence of Crohn's disease-associated adherent-
invasive *E. coli* strain LF82. *PLoS One* **5**, doi:10.1371/journal.pone.0012714 (2010).
- 270 Achtman M, M. G., Zhu P, Wirth T, Diehl I, Kusecek B, Vogler AJ, Wagner DM, Allender
CJ, Easterday WR, Chenal-Francisque V, Worsham P, Thmoson NR, Parkhill J, Lindler
LE, Carniel E, Keim P. Microevolution and history of the plague bacillus, *Yersenia pestis*.
Proc Natl Acad Sci U S A **10**, 17937-17842 (2004).
- 271 Foster JT, B.-S. S., Pearson T, Beckstrom-Sternberg JS, Chain PSG, Roberto FF,
Hinath J, Brettin T, Keim P. Whole-genome-based phylogeny and divergence of the
genus *Brucella*. *J Bacteriol* **191**, 2864-2870 (2009).
- 272 Galloway-Pena P, R. J., Latorre M, Qin X, Murray BE. Genomic and SNP analyses
demonstrate a distant separation of the hospital and community-associated clades of
Enterococcus faecium. *PLoS ONE* **7** (2012).
- 273 Pearson T, G. P., Beckstrom-Sternberg S, Auerbach R, Hornstra H, Tuanyok A, Price
EP, GLass MB, Leadem B, Beckstrom-Sternberg JS, Allan GJ, Foster JT, Wagner DM,
Okinaka RT, Sim SH, Pearson O, Wu Z, Chang J, Kaul R, Hoffmaster AR, Brettin TS,
Robison RA, Mayo M, Gee JE, Tan P, Currie BJ, Keim P. Phylogeographic
reconstruction of a bacterial species with high levels of lateral gene transfer. *BMC Biol* **7**
(2009).
- 274 Lenski RE, W. C., Riley MA. Rates of DNA sequence evolution in experimental
populations of *Escherichia coli* during 20,000 generations. *J Mol Evol* **56**, 498-508
(2003).

- 275 Giron, J. A. Expression of flagella and motility by *Shigella*. *Mol Microbiol* **18**, 63-75 (1995).
- 276 Pupo, G. M., Lan, R. T. & Reeves, P. R. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. *Proc Natl Acad Sci U S A* **97**, 10567-10572 (2000).
- 277 Fricke, W. F. *et al.* Antimicrobial resistance-conferring plasmids with similarity to virulence plasmids from avian pathogenic *Escherichia coli* strains in *Salmonella enterica* serovar Kentucky isolates from poultry. *Appl Environ Microbiol* **75**, 5963-5971 (2009).
- 278 Deodhar, L. P., Saraswathi, K. & Varudkar, A. *Aeromonas* spp. and their association with human diarrheal disease. *Journal of clinical microbiology* **29**, 853-856 (1991).
- 279 Tominaga, A., Lan, R. & Reeves, P. R. Evolutionary changes of the flhDC flagellar master operon in *Shigella* strains. *J Bacteriol* **187**, 4295-4302, doi:10.1128/JB.187.12.4295-4302.2005 (2005).
- 280 Kirby, J. R. Chemotaxis-like regulatory systems: unique roles in diverse bacteria. *Annu Rev Microbiol* **63**, 45-59, doi:10.1146/annurev.micro.091208.073221 (2009).
- 281 Nishiyama, S. *et al.* Thermosensing function of the *Escherichia coli* redox sensor Aer. *J Bacteriol* **192**, 1740-1743, doi:10.1128/JB.01219-09 (2010).
- 282 Yang, Y. S., V. Opposite responses by different chemoreceptors set a tunable preference point in *Escherichia coli* pH taxis. *Mol Microbiol* **86**, 1482-1489 (2012).
- 283 Raterman, E. L. & Welch, R. A. Chemoreceptors of *Escherichia coli* CFT073 Play Redundant Roles in Chemotaxis toward Urine. *PLoS One* **8**, e54133, doi:10.1371/journal.pone.0054133 (2013).
- 284 Gosink, K. K., Buron-Barral, M. C. & Parkinson, J. S. Signaling interactions between the aerotaxis transducer Aer and heterologous chemoreceptors in *Escherichia coli*. *J Bacteriol* **188**, 3487-3493, doi:10.1128/JB.188.10.3487-3493.2006 (2006).
- 285 Alteri, C. J. & Mobley, H. L. T. *Escherichia coli* Physiology and Metabolism Dictates Adaptation to Diverse Host Microenvironments. *Curr Opin Microbiol.* **15**, 3-9 (2012).
- 286 Porcheron, G., Kut, E., Canepa, S., Maurel, M. C. & Schouler, C. Regulation of fructooligosaccharide metabolism in an extra-intestinal pathogenic *Escherichia coli* Strain. *Mol Microbiol* **81**, 717-733 (2011).
- 287 Boin, M. A., Austin, M. J. & Hase, C. C. Chemotaxis in *Vibrio cholerae*. *FEMS Microbiol Lett* **239**, 1-8, doi:10.1016/j.femsle.2004.08.039 (2004).
- 288 Nowrouzian, F. L., Adlerberth, I. & Wold, A. E. Enhanced persistence in the colonic microbiota of *Escherichia coli* strains belonging to phylogenetic group B2: role of virulence factors and adherence to colonic cells. *Microbes Infect* **8**, 834-840 (2006).
- 289 Pickering R, D. P., Jinnah Z, de Ruiter DJ, Churchill SE, Herries AIR, Woodhead JD, Hellstrom JC, Berger LR. *Australopithecus sediba* at 1.977 Ma and implications for the origins of the genus *Homo*. *Science* **333**, 1421-1423 (2011).
- 290 Hou S, L. R., Boudko D, Riley CW, Karatan E, Zimmer M, Ordal GW, Alam M. Myoglobin-like aerotaxis transducers in Archaea and Bacteria. *Nature* **403** (2000).
- 291 Potocka I, T. M., O'Sterias M, Jenal U, Alley MR. Degradation of a *Caulobacter* soluble cytoplasmic chemoreceptor is ClpX dependent. *J Bacteriol* **184**, 6635-6641 (2002).

Appendices

Appendix A

Figure S1. MALS-SEC trace of Tm14_s receptor fragment (residues 107-192)

Figure S2. Electron Density Maps

Figure S3. Pairwise sequence alignments between *E. coli* CheW protein and CheW1

Figure S4. Multiple sequence alignment of full-length CheW protein sequences from *Thermotogae*

Figure S5. Multiple sequence alignment of the signaling domains from 92 MCPs from *Thermotogae* that were assigned to heptad classes.

Figure S6. Multiple sequence alignment of P5 (CheW-like) domains from CheA protein sequences from *Thermotogae*.

Figure S7. A pair of allele-specific repressor mutations between Tsr and *E. coli* CheW

Table S1. The list of CheA, CheW, and MCP protein sequences from *Thermotogae* genomes.

Table S2. Contact site in MCP-CheW and MCP-CheA interactions prospected using 6 Å beta carbon cutoff dependent protein scanning.

Table S3. MCP residues interacting with CheW and P5 revealed by three independent approaches

Table S4. CheW residues interacting with MCPs revealed by three independent approaches

Table S5. CheA residues interacting with MCPs revealed by two independent approaches

Appendix B

Figure S1. Phylogenetic tree of *Escherichia-Shigella*

Figure S2. Presence of chemotaxis genes in *Escherichia/Shigella* genomes

Table S1. *E. coli* genomes used in molecular clock analysis

“Dataset B1.xlsx”

Vita

Aaron D. Fleetwood was born in Albuquerque, New Mexico to Daniel and Betsy Fleetwood. He is the oldest of three brothers: Zachary and Nathan. He was inspired at an early age by his parents to be curious and explore the natural world, both from his father's readings from Jurassic Park by Michael Crichton and countless trips with his mother to the New Mexico Museum of Natural History. These early experiences served as a subconscious force that steered him, eventually, toward the natural and life sciences. After moving from New Mexico to Middle Tennessee, Aaron attended and graduated from Brentwood High School before applying to Vanderbilt University, where he was admitted to the College of Arts and Sciences. While at Vanderbilt, Aaron double-majored in Economics and East Asian Studies, narrowly choosing an international business track over the biological sciences. He also took several programming courses and conducted summer undergraduate computing research. While at Vanderbilt, he met the love of his life, Ellen Messenger. After graduating in 2008, he soon began to realize that he wanted to pursue the interface between biology, computing, and medicine. In no small part due to Ellen's encouragement and support, he was accepted to the University of Tennessee Knoxville Graduate Program in Genome Science and Technology, where his passion for biology and hands on computing research experience found a perfect match. He joined Dr. Igor Zhulin's lab, where he has truly found an ideal career: using comparative genomics to answer biological questions. Aaron and Ellen (now a DVM) are now married and live with their two cats and a guinea pig. In the future, Aaron hopes to use his comparative genomics methods in larger organisms (animals and humans), in order to inform and improve therapeutic strategies, genetic disease detection, and preventative medicine. He will also continue to pursue a more complete understanding of motile bacteria and the chemoreceptors that drive them.